

June I-5, L’Escala, Girona

# Proceedings of the 6th International Workshop on Compositional Data Analysis 

S. Thió-Henestrosa and J.A. Martín Fernández (Editors)

## Sponsored by:



Promoted by:
iasc

## Welcome to CoDawork 2015

CoDaWork 2015, the 6th international Workshop on Compositional Data analysis, offers a forum of discussion for people concerned with the statistical treatment and modelling of compositional data or other constrained data sets, and the interpretation of models or applications involving them. The primary goal of the workshop is to identify important potential lines of future research and gain insight as to how they might be tackled.

CoDaWork 2015 intends to bring together specialist researchers, data analysts, postgraduate students, as well as those with a general interest in the field, to summarize and share their contributions and recent developments.

Compositional data (CoDa) are typically defined as vectors of positive components and constant sum, usually $100 \%$ or I. These conditions render most classical statistical techniques incoherent on compositions, as they were devised for unbounded real vectors. However, there are many more types of data having the same limitations: as soon as the variables of a data set show the relative importance of some parts of a whole, data should be considered compositional. Examples of disguised compositions are data presented in ppm, ppb, molarities, or any other concentration units. Aitchison introduced the log-ratio approach to analyse CoDa back in the eighties. His solution was based on transforming the data vector with some log-ratio transformations, and applying classical techniques to the scores so obtained. This became the foundation of modern CoDa analysis, nowadays based on an own geometric structure for the simplex, an appropriate representation of the sample space of CoDa . The validity of these considerations is not restricted to CoDa: there are many more data sets which sample space is not obeying the rules of real numbers, or that can be given an own, alternative, meaningful geometric structure. Examples abound in the natural and social sciences: vectors of positive amounts, functional data, spherical data, ordered variables, etc. CoDa analysis insights may be of good use to scientists working with these data sets, and vice versa.

Practitioners interested in CoDA can find in the website: www.compositionaldata.com a forum for the exchange of information, material and ideas. Also a free software can be found there, the Compositional Data Package (CoDaPack), which implements presently the most elementary the compositional statistical methods. It can be downloadable from: http://ima.udg.edu/codapack. This software has been developed and is maintained by the research group pn compositional data analysis of the Universitat de Girona, and is oriented to users coming from the applied sciences, with no extensive background in using various computer packages.

In www.compositionaldata.com also R Packages can be found: zCompositions, compositions, robCompositions.

This publication is hosted at the University of Girona Digital Repository:
http://dugi-doc.udg.edu/handle/I0256/I50
To cite the whole document, according the citation style ISO 690:1987 (UNE 50-10494), use:

Proceedings of the 6th International Workshop on Compositional Data Analysis: Girona, I-7 de juny de 2015. Girona: Departament d'Informàtica, Matemàtica Aplicada i Estadística, Universitat de Girona, 2015. 3I7p. ISBN 978-84-8458-45I-3.

To cite specific proceeding, according the same style, use, for example:
Egozcue, J. J.; Pawlowsky-Glahn, V. Changing the reference measure in the simplex and its weighting effects. Proceedings of the 6th International Workshop on Compositional Data Analysis: Girona, I-5 de juny de 2015. Girona: Universitat de Girona, 2015, P. 7685.

We acknowledge work of all organizers and the support of our hosts and sponsors and promotors, and particularly the Consell Social of the University of Girona, the Spanish Ministerio de economía y competitividad (Statistical MEthods in resTRICted spaceS (METRICS). Ref. MTM20I2-33236), the Catalan Agència de gestió d’ajuts universitaris i de recerca of the Generalitat de Catalunya (Compositional and Spatial Analysis (COSDA). Ref: 2014SGR55I), the Austrian Statistical Society (ÖSG), the Statistical Modelling Society (SMS), the International Association for Mathematical Geology (IAMG), the L'Escala city council, and the International Association for Statistical Computing (IASC).

We wish you an useful, inspiring workshop and a nice stay in L'Escala.
L’Escala, June I, 2015

## Contents

I J. BEAR and D. BILLHEIMER: A Logistic Normal Mixture Model Allowing Essential Zeros

2 J. BERGMAN and A. LINDAHL: Optimising Archaeologic Ceramics XRF Analyses
3 M.S. BLONDES, M.A. ENGLE and N.J. GEBOY: Advances in integrating isotopic data with compositional data analysis: applications for deep formation brine geochemistry

4 K. G. VAN DEN BOOGAART, M. LEHMANN and R. TOLOSANA-DELGADO: linear models for the conditional distribution

5 A. BUCCIANTI, B. NISI, and B. RACO: From univariate background (baseline) values towards the concept of compositional background (baseline) values

6 G. COENDERS, B. FERRER-ROSELL, G. MATEU-FIGUERAS, and V. PAWLOWSKYGLAHN: MANOVA of Compositional Data with a Total

7 J. J. EGOZCUE and V. PAWLOWSKY-GLAHN: Changing the reference measure in the simplex and its weighting effects

8 P. H. C. EILERS, M. DE ROOIJ: Semi-parametric density estimation with TrioScale
9 I. GALVÁN, J. GRAFFELMAN and C. BARCELÓ: A compositional approach to allele sharing analysis

IO G. B. GLOOR, A. D. FERNANDES, J. M. MACKLAIM, and M. VU: Compositional uncertainty in high-throughput sequencing data analysis should not be ignored.

II J. GRAFFELMAN and R. DE WIT: Microbial Denitrification and the HardyWeinberg law

12 J J. GRAFFELMAN, J. J. EGOZCUE and M.I. ORTEGO: On the asymptotic distribution of proportions of multinomial count data

13 J. GRAFFELMAN and I. GALVÁN: An application of the isometric log-ratio transformation in relatedness research

14 J. GRAFFELMAN and R. TOLOSANA-DELGADO: Biplots for Compositional Regression

I5 E.C. GRUNSKY and B.A. KJARSGAARD: Recognizing and Validating Structural Processes in Geochemical Data

16 M. K. HOSSEINPOOR, H. MOINI and F. M. TORAB: Detecting and mapping compositional global outliers to identify mineral exploration targets, Case study: Khusf district, East of Iran

I7 E. JARAUTA-BRAGULAT and J.J. EGOZCUE: Space-time compositional models: an introduction to simplicial partial differential operators

18 G.S. MONTI, G. MATEU-FIGUERAS, V. PAWLOWSKY-GLAHN and J.J. EGOZCUE: Shifted-Dirichlet Regression vs Simplicial Regression: a comparison

19 M.I. ORTEGO, J.J. EGOZCUE: Bayesian estimation of the orthogonal decomposition of a contingency table

20 R. ORTELLS, JJ. EGOZCUE, M.I. ORTEGO and A. GAROLA: Relationship between popularity of key words in the Google browser and the evolution of worldwide financial indexes

21 J. PALAREA-ALBALADEJO and S. F. M. CHASTIN: Physical Activity and Health Outcomes Analysed using Compositional Methods

22 V. PAWLOWSKY-GLAHN, T. MONREAL-PAWLOWSKY and J.J. EGOZCUE: Representation of species composition

23 P.A. ROESER, S.O. FRANZ and T. LITT: Insights on the bio-geochemical processes that control sedimentation in Lake Iznik based on compositional data analysis

24 R. TOLOSANA-DELGADO, K. G. VAN DEN BOOGAART, E. FISEROVA, K. HRON and I. DUNKL: Joint compositional calibration: a geochronological example

25 H. WANG, L. HUANG, L. SHANGGUAN and S. WANG: Variable Selection and Estimation for Regression Models with Compositional Data Predictors

26 N. C. WARMAN: Multi element geochemical modelling for mine planning: Case study from an epithermal gold deposit

27 T. YOSHIDA and M. TSUTSUMI: Regression model for compositional data using a spatial econometric approach

28 Z. ZIEMBIK, A. DOLHANCZUK-RÓDKA: Application of compositional data analysis in biomonitoring of atmospheric dust precipitation

# A Logistic Normal Mixture Model Allowing Essential Zeros 

John Bear and Dean Billheimer<br>Statistical Consulting Lab, University of Arizona, U.S.A.<br>jbear@email.arizona.edu, dean.billheimer@arizona.edu

June 16, 2015


#### Abstract

The usual candidate distributions for modeling compositions, the Dirichlet and the logistic normal distribution, do not include zero components in their support. Methods have been developed and refined for dealing with zeros that are rounded, or due to a value being below a detection level, Palarea-Albaladejo et al. (2014); Pawlowsky-Glahn and Buccianti (2011); Martin-Fernandez et al. (2011); Palarea-Albaladejo and MartinFernandez (2008); Fry et al. (2000). Methods have also been developed for zeros in compositions arising from count data, Billheimer et al. (2001); Bacon-Shone (2008); Daunis-i-Estadella et al. (2008).

However, essential zeros, cases where a component is truly absent, in continuous compositions are still a problem. Aitchison and Kay (2003) proposed an approach for extending the logistic normal distribution to model essential zeros. They posited a mixture of additive logistic normal distributions of different dimension, related by common parameters. They discussed possible ways of estimating parameters but did not derive estimators for location or dispersion parameters. We continue their approach, and by imposing an additional constraint, develop a likelihood, and show ways of estimating parameters for location and dispersion. The proposed likelihood, conditional on parameters for the probability of zeros, is a mixture of additive logistic normal distributions of different dimensions whose location and dispersion parameters are projections of a common location or dispersion parameter. For some simple special cases, we contrast the relative efficiency of different location estimators.


## 1 Introduction

An essential zero in compositional data is a zero component which is not caused by rounding or some other difficulty in measurement, but rather, is genuinely believed to be zero. This is fundamentally a different problem than that addressed by recent work on rounded zeros, or below-detection level zeros, such as in Palarea-Albaladejo et al. (2014), Martin-Fernandez et al. (2011), Palarea-Albaladejo and Martin-Fernandez (2008), Fry et al. (2000), Aitchison
(1986). Although there are workable Bayesian approaches to zeros in compositions from count data, Billheimer et al. (2001); Bacon-Shone (2008); Daunis-i-Estadella et al. (2008), essential zeros in continuous compositions still present a problem.

We develop an approach proposed by Aitchison and Kay (2003) to extend the logistic normal distribution to accommodate essential zeros. Aitchison (1986) and Aitchison and Kay (2003) note that a key feature compositional data is that ratios of the components contain all pertinent information about the composition. Essential zeros complicate this feature in that they contain no information about the other components of the composition. In addition, an observation containing an essential zero is at the boundary of the simplex and is a composition of smaller dimension.

## 2 Previous Work

In addition to the work mentioned above, there have been other approaches to zeros in compositions. Work by Butler and Glasbey (2008) mapped a latent Gaussian variable to a composition, but seems only to work for two and three place compositions. In contrast, Leininger et al. (2013) have developed a more practical treatment of compositions as coming from a latent Gaussian random variable where the compositional component is zero when the latent Gaussian component is less than or equal to zero. They develop a hierarchical power model with the transformation $Y_{k}=\frac{\left(\max \left(0, Z_{k}\right)\right)^{\gamma}}{1+\sum_{k^{\prime}=1}^{\left(\max \left(0, z_{k^{\prime}}\right)^{\gamma}\right.}}$ where $z_{k}$ is the $k^{\text {th }}$ normal component and $Y_{k}$ is the corresponding compositional component. To estimate parameters they use MCMC. One limitation of their approach is also a limitation of ours: we require one component of the composition to be strictly positive.

Work by Stewart and Field (2011) uses a multiplicative logistic normal mixture model that allows them to consider the univariate log odds for the $i^{\text {th }}$ component to be normally distributed where the $i^{\text {th }}$ component is not zero. It works well for their applications, in particular regression, but does not capture covariance easily.

Scealy and Welsh (2011) transform compositions into directional data on the hypersphere, and develop a regression model using the Kent distribution, Kent (1982), which tolerates zeros provided there are not very many.

Our goal here is to extend the additive logistic normal distribution to handle essential zeros for continuous data.

## 3 Definitions (from Aitchison, 1986)

## Definition: If

$$
\mathbf{x}=\left(x_{1}, x_{2}, \ldots, x_{d}, x_{D}\right)^{T}, \text { and } x_{i} \geq 0, \text { and } \sum_{i=1}^{D} x_{i}=1
$$

then $\mathbf{x}$ is called a composition.
Definition: The d-dimensional simplex embedded in D-dimensional real space is the set defined by

$$
\mathscr{S}^{d}=\left\{\left(x_{1}, \ldots, x_{D}\right): x_{1}>0, \ldots, x_{D}>0 ; \sum_{i=1}^{D} x_{i}=1\right\}
$$

Convention: $d=D-1$.
Definition:

$$
\text { If } \mathbf{x}=\left(x_{1}, x_{2}, \ldots, x_{d}, x_{D}\right)^{T} \text {, then } \mathbf{x}_{-D}=\left(x_{1}, x_{2}, \ldots, x_{d}\right)^{T}
$$

Definition: The additive logratio transformation, $\phi$, is defined as follows:

$$
\begin{aligned}
\phi: \mathscr{S}^{d} & \rightarrow \mathbb{R}^{d} \\
\quad \mathbf{x} & \mapsto \mathbf{y}=\left(\log \left(x_{1} / x_{D}\right), \log \left(x_{2} / x_{D}\right), \ldots, \log \left(x_{d} / x_{D}\right)\right)^{T} .
\end{aligned}
$$

We define the shorthand:

$$
\log \left(\mathbf{x}_{-D} / x_{D}\right)=\left(\log \left(x_{1} / x_{D}\right), \log \left(x_{2} / x_{D}\right), \ldots, \log \left(x_{d} / x_{D}\right)\right)^{T}
$$

Since $\phi$ is one-to-one, its inverse exists. It is called the logistic transformation, $\phi^{-1}$, defined as:

$$
\begin{aligned}
x_{i} & =\exp \left(y_{i}\right) /\left\{\exp \left(y_{1}\right)+\cdots+\exp \left(y_{d}\right)+1\right\} \quad(i=1, \ldots, d), \\
x_{D} & =1-x_{1}-x_{2}-\cdots-x_{d} \\
& =1 /\left\{\exp \left(y_{1}\right)+\cdots+\exp \left(y_{d}\right)+1\right\} .
\end{aligned}
$$

## 4 Simplifying Assumption

In this section we outline our method for building a mixture distribution for dealing with compositions containing essential zeros, but leave most of the details about the weights for later. A key simplifying assumtion we make throughout is that one of the components of the composition, the $D^{t h}$ component, is never zero. We allow zeros anywhere else but not in the last component.

Let $\mathbf{x}=\left(x_{1}, x_{2}, \ldots, x_{d}, x_{D}\right)^{T}$ be a composition with $x_{i}<1$ for all $i \in\{1,2, \ldots, d, D\}$ and $x_{D}>0$. For $i \in\{1,2, \ldots, d\}$, consider two possibilities. Either $x_{i}=0$ or $x_{i}>0$. Let $W=\left\{i: i \in\{1,2, \ldots, d\}, x_{i}>0\right\}$. That is, $W$ is the set of indices for the components of $\mathbf{x}$ (other than $x_{D}$ ) which are nonzero (positive). For any given composition $\mathbf{x}, W$ is the set of all the indices of the nonzero components of $\mathbf{x}$. There are $2^{d}-1$ possible sets $W$. There are $2^{d}-1$ and not $2^{d}$ because $W$ cannot be empty. If $W$ were empty that would require that $x_{D}=1$ in order for $\mathbf{x}$ to be a composition, but we have already said we require all $x_{i}<1$ including $x_{D}$. Each pattern of zeros corresponds to a different set $W$. We index them as $W_{\ell}$ with $\ell \in\left\{1,2, \ldots, 2^{d}-1\right\}$. They are elements of the power set,
$W_{\ell} \in \mathscr{P}(\{1,2, \ldots, d\})$. Sometimes we refer to these sets with incidence vectors where the $i^{t h}$ component $V_{W_{\ell i}}=1 \Longleftrightarrow x_{i}>0$ and $V_{W_{\ell i}}=0 \Longleftrightarrow x_{i}=0$.

Each $W_{\ell}$ has some probability of occurrence, $P\left(W_{\ell}\right)$. The probabilities must sum to one,

$$
\sum_{\ell=1}^{2^{d}-1} P\left(W_{\ell}\right)=1
$$

We use the probabilities $P\left(W_{\ell}\right)$ as the weights in a mixture distribution. For the other distributions making up our mixture, we use logistic normal distributions $\mathscr{L}\left(\mathbf{x} ; \boldsymbol{\mu}_{W_{\ell}}, \boldsymbol{\Omega}_{W_{\ell}}\right)$ derived from a single parent logistic normal distribution $\mathscr{L}(\mathbf{x} ; \boldsymbol{\mu}, \boldsymbol{\Omega})$. They are in fact projections from the parent. We will call the distributions derived from the parent distribution subdistributions once we define them. So the mixture distribution will be denoted as follows, once we define a few more terms,

$$
g(\mathbf{x} ; \boldsymbol{\mu}, \boldsymbol{\Omega})=\sum_{\ell=1}^{2^{d}-1} P\left(W_{\ell}\right) \mathscr{L}\left(\mathbf{x} ; \boldsymbol{\mu}_{W_{\ell}}, \boldsymbol{\Omega}_{W_{\ell}}\right)
$$

In the parent distribution, $\mathscr{L}(\mathbf{x} ; \boldsymbol{\mu}, \boldsymbol{\Omega}), \boldsymbol{\mu}$ is a $d$-place location parameter vector, $\boldsymbol{\mu} \in R^{d}$, and $\boldsymbol{\Omega}$ is a $d \times d$ positive definite dispersion matrix. To ease the discussion we will refer to $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ as mean vector and variance-covariance matrix respectively, although they are not moments of the distribution. For the distributions derived from the logistic normal parent distribution, the parameters $\boldsymbol{\mu}_{W_{\ell}}$ and $\boldsymbol{\Omega}_{W_{\ell}}$ are defined in terms of the parameters $\boldsymbol{\mu}$, and $\boldsymbol{\Omega}$, and the set of indices of nonzero components of $\mathbf{x}, W_{\ell}$, and a selection matrix $\mathbf{B}_{W_{\ell}}$.

Next we define our selection matrix $\mathbf{B}_{W_{\ell}} . W_{\ell} \subset\{1,2,3, \ldots, d\}$ is a nonempty set of indices (of the nonzero components of $\mathbf{x}$ ), and without loss of generality we can order the indices from least to greatest:

$$
W_{\ell}=\left\{j_{1}, j_{2}, \ldots, j_{J}\right\} \text { where } 0<j_{1}<j_{2}<\ldots<j_{J} \leq d .
$$

Now we define our $J \times d$ selection matrix, $\mathbf{B}_{W_{\ell}}=\left[B_{i, m}\right]$.

$$
\text { For } i \in\{1,2, \ldots, J\} \text {, and } m \in\{1,2, \ldots, d\} \text {, with } W_{\ell}=\left\{j_{1}, j_{2}, \ldots, j_{J}\right\} \text {, }
$$

we define the elements of $\left[B_{i, m}\right]$ to be

$$
B_{i, j_{i}}=1 \text { and } B_{i, m \neq j_{i}}=0 .
$$

For example, let $\mathbf{x}=(.2,0, .3,0, .25, .25)$, a 6 -place composition, with $x_{6}>0$. The set of nonzero indices is $W_{\ell}=\{1,3,5\}$, and the selection matrix is

$$
\mathbf{B}_{W_{\ell}}=\mathbf{B}_{(1,3,5)}=\left(\begin{array}{ccccc}
1 & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 1
\end{array}\right)
$$

Now it is easy to define $\boldsymbol{\mu}_{W_{\ell}}$ and $\boldsymbol{\Omega}_{W_{\ell}}$. We define:

$$
\begin{gathered}
\boldsymbol{\mu}_{W_{\ell}}=\left(\mathbf{B}_{W_{\ell}}\right)(\boldsymbol{\mu}) . \\
\boldsymbol{\Omega}_{W_{\ell}}=\left(\mathbf{B}_{W_{\ell}}\right)(\boldsymbol{\Omega})\left(\mathbf{B}_{W_{\ell}}^{T}\right) .
\end{gathered}
$$

With this structure in place, the mixture distribution can now be more fully specified.

$$
g(\mathbf{x} ; \boldsymbol{\mu}, \Omega)=\sum_{\ell=1}^{2^{d}-1} P\left(W_{\ell}\right) \mathscr{L}^{\left\|W_{\ell}\right\|}\left(\mathbf{x} ; \boldsymbol{\mu}_{W_{\ell}}, \boldsymbol{\Omega}_{W_{\ell}}\right) \text { where }
$$

- \| $W_{\ell} \|$ refers to the cardinality of the set $W_{\ell}$.
- $\sum P\left(W_{\ell}\right)=1$.
- $\mu$ is a $d$-place vector in $R^{d}$.
- $\boldsymbol{\mu}_{W_{\ell}}$ is a subvector of $\boldsymbol{\mu}$ corresponding to the $W_{\ell}$ pattern of zeros.
- $\boldsymbol{\Omega}_{W_{\ell}}$ is a submatrix of a $d \times d$ positive definite covariance matrix corresponding to the $W_{\ell}$ pattern of zeros.


### 4.1 Multivariate Normal Foundation

In our approach there is a tight correspondence between the $y_{i}$ parts of a multivariate normal vector and the $x_{i}$ parts of a composition, possibly one containing essential zeros.

$$
\begin{gathered}
\qquad y_{i}=\log \left(x_{i} / x_{D}\right) \text { for } i=1,2,3, \ldots, d . \\
\text { Composition: } \mathbf{x}=\left(x_{1}, x_{2}, x_{3}, \ldots, x_{d}, x_{D}\right)^{T} \\
\text { Logratio transformed vector: } \mathbf{y}=\log \left(\mathbf{x}_{-D} / x_{D}\right)=\left(y_{1}, y_{2}, y_{3}, \ldots, y_{d}, \odot\right)^{T}
\end{gathered}
$$

When there is an essential zero in the composition in one of the $x_{i}$ slots, e.g., in $x_{2}$, we use $\odot$ as a placeholder so things line up, for example:

$$
\begin{aligned}
\text { Composition: } \mathbf{x} & =\left(x_{1}, 0, x_{3}, \ldots, x_{d}, x_{D}\right)^{T} \\
\text { Logratio transformed subvector: } \mathbf{y}=\log \left(\mathbf{x}_{-\{2, D\}} / x_{D}\right) & =\left(y_{1}, \odot, y_{3}, \ldots, y_{d}, \odot\right)^{T}
\end{aligned}
$$

When we have an essential zero in the $i^{\text {th }}$ component of the composition, we have a selected subvector of the $y$ 's which does not contain an element corresponding to $y_{i}$. The requirement that $x_{D}>0$ is what allows us to maintain this strict correspondence between $x_{i}$ and $y_{i}$.

Next we turn to defining an extension to Aitchison's logistic normal distribution. Recall Aitchison defined the logistic normal distribution (without zeros) as:

Let $\mathbf{x}=\left(x_{1}, x_{2}, x_{3}, \ldots, x_{d}, x_{D}\right)^{T}$ be a composition.
Let $\mathbf{y}=\log \left(\mathbf{x}_{-D} / x_{D}\right)=\left(y_{1}, y_{2}, y_{3}, \ldots, y_{d}\right)^{T}$.
Then we define $\mathbf{x}$ to have a logistic normal distribution with location $\boldsymbol{\mu}$ and dispersion $\boldsymbol{\Omega}$, written $\mathbf{x} \sim \mathscr{L}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, if $\mathbf{y} \sim \mathscr{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$.

Note also that if $\mathbf{y}$ is distributed as a multivariate normal, $\mathbf{y} \sim \mathscr{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, and $\mathbf{B}$ is a selection matrix as mentioned in Section 4, then By is also multivariate normal with distribution:

$$
\mathbf{B y} \sim \mathscr{N}\left(\mathbf{B} \boldsymbol{\mu}, \mathbf{B} \boldsymbol{\Omega} \mathbf{B}^{T}\right) .
$$

And finally, recall that also in Section 4 we used $W_{\ell}$ to represent the set of indices of the nonzero components of $\mathbf{x} . \mathbf{B}_{W_{\ell}}$ is the selection matrix that selects those components when we perform these matrix multiplications:

$$
\begin{gathered}
\boldsymbol{\mu}_{W_{\ell}}=\left(\mathbf{B}_{W_{\ell}}\right)(\boldsymbol{\mu}) . \\
\boldsymbol{\Omega}_{W_{\ell}}=\left(\mathbf{B}_{W_{\ell}}\right)(\boldsymbol{\Omega})\left(\mathbf{B}_{W_{\ell}}^{T}\right) .
\end{gathered}
$$

With all these definitions in place we are now in a position to define a logistic normal distribution with essential zeros.
Definition:
Let $\mathbf{x}=\left(x_{1}, x_{2}, x_{3}, \ldots, x_{d}, x_{D}\right)^{T}$ be a composition with $\mathbf{x}_{D}>0$.
Let $W_{\ell}=\left\{i_{1}, i_{2}, \ldots, i_{r}\right\} \subset\{1,2, \ldots, d\}$ be a nonempty set of indices of nonzero components of $\mathbf{x}$.
Let $\mathbf{B}_{W_{\ell}}$ be the corresponding selection matrix.
Let $\mathbf{y}=\log \left(\mathbf{B}_{W_{\ell}} \mathbf{x}_{-D} / x_{D}\right)=\left(y_{i_{1}}, y_{i_{2}}, \ldots, y_{i_{r}}\right)^{T}=\tilde{\phi}\left(\mathbf{x}, W_{\ell}, D\right)$ be the logratios of the nonzero components of $\mathbf{x}$.

If for every set $W_{\ell}$ of indices of nonzero components of $\mathbf{x}$, we have $\mathbf{y} \sim \mathscr{N}\left(\boldsymbol{\mu}_{W_{\ell}}, \boldsymbol{\Omega}_{W_{\ell}}\right)$, then $\mathbf{x}$ has a logistic normal distribution with essential zeros, written $\mathbf{x} \sim \mathscr{L}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, with probability density function

$$
g(\mathbf{x} ; \boldsymbol{\mu}, \Omega)=\sum_{\ell=1}^{2^{d}-1} P\left(W_{\ell}\right) \mathscr{L}^{\left\|W_{\ell}\right\|}\left(\mathbf{x} ; \boldsymbol{\mu}_{W_{\ell}}, \boldsymbol{\Omega}_{W_{\ell}}\right)
$$

where

$$
\sum P\left(W_{\ell}\right)=1
$$

$\boldsymbol{\mu}$ is a $d$-place vector in $R^{d}$.
$\boldsymbol{\mu}_{W_{\ell}}$ is a subvector of $\boldsymbol{\mu}$ corresponding to the $W_{\ell}$ pattern of zeros.
$\boldsymbol{\Omega}$ is a $d \times d$ positive definite covariance matrix.
$\boldsymbol{\Omega}_{W_{\ell}}$ is a square submatrix of $\boldsymbol{\Omega}$, corresponding to the $W_{\ell}$ pattern of zeros.
Now we extend the notation for the inverse of the logratio transformation, $\phi^{-1}$ from Aitchison (1986). We use the new symbols, $\tilde{\phi}$ and $\tilde{\phi}^{-1}$. We define them in terms of $W$ and $D$, the maximum index. Let $W \subset\{1,2, \ldots, d\}$ be a pattern of zeros, i.e., a set of indices of nonzero components of $\mathbf{x}$ and denote them: $W=\left\{i_{1}, i_{2}, \ldots, i_{r}\right\}$, and let $j \in\{1,2, \ldots, d, D\}$. Now we define $\tilde{\phi}^{-1}$.

$$
\begin{align*}
& \tilde{\phi}^{-1}(\mathbf{y}, W, D)=\left(x_{1}, \ldots, x_{d}, x_{D}\right)^{T} \text { where }  \tag{1}\\
& x_{j}= \begin{cases}\exp \left(y_{j}\right) /\left\{\exp \left(y_{i_{1}}\right)+\exp \left(y_{i_{2}}\right)+\ldots+\exp \left(y_{i_{r}}\right)+1\right\} & \text { if } j \in W \\
0 & \text { if } j \notin W \& j \in\{1, \ldots, d\} \\
1 /\left\{\exp \left(y_{i_{1}}\right)+\exp \left(y_{i_{2}}\right)+\ldots+\exp \left(y_{i_{r}}\right)+1\right\} & \text { if } j=D\end{cases}
\end{align*}
$$

For the case where $W=\{1,2, \ldots, d\}$ the composition $\mathbf{x}=\left(x_{1}, x_{2}, \ldots, x_{D}\right)^{T}$ has the additive logistic normal distribution, $\mathscr{L}^{d}(\boldsymbol{\mu}, \boldsymbol{\Omega})$.

### 4.2 Common Expectations and Variances

The definition of $\tilde{\phi}^{-1}$ enables compositions from different subdistributions to be used to estimate parameters of their shared parent distribution. Let $\mathbf{x}_{1}=\left(x_{11}, x_{21}, \ldots, x_{D 1}\right)^{T}$, and let $\mathbf{x}_{2}=\left(x_{12}, x_{22}, \ldots, x_{D 2}\right)^{T}$ with

$$
\begin{align*}
& \mathbf{x}_{1} \sim \mathscr{L}^{\left|W_{1}\right|}\left(\boldsymbol{\mu}_{W_{1}}, \boldsymbol{\Omega}_{W_{1}}\right), \text { and }  \tag{2}\\
& \mathbf{x}_{2} \sim \mathscr{L}^{\left|W_{2}\right|}\left(\boldsymbol{\mu}_{W_{2}}, \boldsymbol{\Omega}_{W_{2}}\right) . \tag{3}
\end{align*}
$$

The two sets of nonzero indices, $W_{1}, W_{2}$ need not have any elements in common, nor do they need to have the same number of elements, though $\mathbf{x}_{1}$ and $\mathbf{x}_{2}$ both have $D$ elements. Suppose they have an index, $m$, in common: $m \in W_{1} \cap W_{2}$. By properties of the logistic normal distribution (Aitchison (1986), p. 116), and the definition of $\tilde{\phi}^{-1}$ in Equation 1 we have:

$$
\begin{equation*}
E \log \left(x_{m 1} / x_{D 1}\right)=E y_{m}=\mu_{m}=E y_{m}=E \log \left(x_{m 2} / x_{D 2}\right) \tag{4}
\end{equation*}
$$

And similarly,

$$
\begin{equation*}
\operatorname{Var}\left[\log \left(x_{m 1} / x_{D 1}\right)\right]=\operatorname{Var}\left[y_{m}\right]=\sigma_{m}^{2}, \text { and } \tag{5}
\end{equation*}
$$

$$
\begin{equation*}
\operatorname{Var}\left[\log \left(x_{m 2} / x_{D 2}\right)\right]=\operatorname{Var}\left[y_{m}\right]=\sigma_{m}^{2} \tag{6}
\end{equation*}
$$

Thus, compositions from different subdistributions of the same logistic normal distribution can be used to estimate the parameters of their shared parent distribution.

## 5 Data Blocks

Now that we have a correspondence between multivariate normal variables and compositions with zeros, we could derive a density function using the standard formula for transformed variables, analogous to Aitchison (1986), chapter 6. However, for estimating parameters it is more convenient to work in the space of the transformed variables (multivariate normal projections).

Here we apply the techniques and notation of block matrices and matrix calculus to do some preparation in order to build a likelihood and attack the problem of finding estimators for the parameters. We discuss two sets of estimators, a general maximum likelihood estimator, and a simpler pair of estimators reminiscent of method of moments estimators.

### 5.1 Block Matrices of Compositions

We write a collection of compositional data with zeros, $\mathbf{X}$, as a column of blocks of compositions where each block, $\mathbf{X}_{\ell}$, has a particular pattern of zeros throughout. That is, for a particular block, $\mathbf{X}_{\ell}$, and $i \in\{1,2, \ldots, d\}$, the $i^{t h}$ column of $\mathbf{X}_{\ell}$ is either all positive, or all zero.

$$
\text { Let } \underset{n \times D}{\mathbf{X}}=\left[\begin{array}{c}
\mathbf{X}_{1}  \tag{7}\\
\mathbf{X}_{2} \\
\vdots \\
\mathbf{X}_{b}
\end{array}\right]
$$

The dimensions of the blocks are: $\underset{r_{1} \times D}{\mathbf{X}_{1}}, \underset{r_{2} \times D}{\mathbf{X}_{2}}, \ldots, \underset{r_{b} \times D}{\mathbf{X}_{b}}$ and the sum of their vertical dimensions is $r_{1}+r_{2}+\ldots+r_{b}=n$, where $n$ is the number of data points.

We use $\ell$ to indicate a block, and $t$ to indicate a composition (row) in that block. Next we define the patterns of zeros in each block. Here $i \in\{1,2, \ldots, D\}$. For $\ell \in\{1,2, \ldots, b\}$, let $W_{\ell} \subset\{1,2, \ldots, d\}$ be the set of indices of strictly positive components of $\mathbf{X}_{\ell}$.

$$
\text { For } \ell \in\{1,2, \ldots, b\}, \mathbf{X}_{\ell}=\left[x_{t i}\right] \text { where }\left\{\begin{array}{l}
x_{t i}>0 \text { if } i=D  \tag{8}\\
x_{t i}>0 \text { if } i \in W_{\ell} \\
x_{t i}=0 \text { if } i \notin W_{\ell} \text { and } i \neq D
\end{array}\right.
$$

Each row of $\mathbf{X}_{\ell}$ is required to be a composition. For each row $t$, where $1 \leq t \leq r_{\ell}$,

$$
\sum_{i=1}^{D} x_{t i}=1
$$

Since each $\mathbf{X}_{\ell}$ consists entirely of compositions, the overall matrix $\mathbf{X}$ does too.

### 5.2 Transformations - Ratios and Logratios

First we define the logratio transformation, $\phi$, for the case where there are no zeros, i.e., $x_{i}>0, \forall i \in\{1,2, \ldots, D\}$.

$$
\begin{align*}
& \phi: \mathscr{S}^{d} \rightarrow \mathbb{R}^{d}  \tag{9}\\
& \mathbf{x} \mapsto \mathbf{y}=\left(\log \left(x_{1} / x_{D}\right), \log \left(x_{2} / x_{D}\right), \ldots, \log \left(x_{d} / x_{D}\right)\right)^{T}, \text { and }  \tag{10}\\
& \mathbf{x}^{T} \mapsto \mathbf{y}^{T}=\left(\log \left(x_{1} / x_{D}\right), \log \left(x_{2} / x_{D}\right), \ldots, \log \left(x_{d} / x_{D}\right)\right) . \tag{11}
\end{align*}
$$

Next we extend $\phi$ to $\tilde{\phi}$ for a block matrix of compositions, $\underset{r_{\ell} \times D}{\mathbf{X}}$ which may contain zeros. We do this by defining a selection matrix $\mathbf{B}_{W_{\ell}}$ corresponding to set $W_{\ell}$. We still have $W_{\ell} \subset\{1,2,3, \ldots, d\}$ being a nonempty set of indices of the nonzero components of $\mathbf{x}$, and without loss of generality we can order the indices from least to greatest:

$$
\begin{equation*}
W_{\ell}=\left\{j_{1}, j_{2}, \ldots, j_{J}\right\} \text { where } 0<j_{1}<j_{2}<\ldots<j_{J}<D . \tag{12}
\end{equation*}
$$

Now we define our $(J+1) \times D$ selection matrix, $\mathbf{B}_{W_{\ell}}=\left[B_{p, m}\right]$. We use $J+1$ here because we construct the selection matrix so that the final, $D^{\text {th }}$, component of the data is always selected. This is slightly different than before. Previously we constructed $B$ to conform to the parameters $\boldsymbol{\mu}(d \times 1)$ and $\boldsymbol{\Omega}(d \times d)$.

For $p \in\{1,2, \ldots, J+1\}$, and $m \in\{1,2, \ldots, D\}$, with $W_{\ell}=\left\{j_{1}, j_{2}, \ldots, j_{J}\right\}$,
we define the elements of $\left[B_{p, m}\right]$ to be $B_{p, j_{p}}=1$ and $B_{p, m \neq j_{p}}=0$.
$\mathbf{X}_{\ell} \mathbf{B}_{W_{\ell}}^{T}$ is a matrix where each row vector is a composition without zeros.

$$
\begin{gather*}
\mathbf{X}_{\ell} \mathbf{B}_{W_{\ell}}^{T}=\left[\begin{array}{c}
\mathbf{x}_{1}^{T} \\
\mathbf{x}_{2}^{T} \\
\vdots \\
\mathbf{x}_{r_{\ell}^{T}}
\end{array}\right] \mathbf{B}_{W_{\ell}}^{T}=\left[\begin{array}{cccc}
x_{11} & x_{12} & \ldots & x_{1(J+1)} \\
x_{21} & x_{22} & \ldots & x_{2(J+1)} \\
\vdots & \vdots & & \vdots \\
x_{r_{\ell} 1} & x_{r_{\ell} 2} & \ldots & x_{r_{\ell}(J+1)}
\end{array}\right] .  \tag{15}\\
\text { We define } \tilde{\phi}\left(\mathbf{X}_{\ell}, W_{\ell}, D\right)=\phi\left(\mathbf{X}_{\ell} \mathbf{B}_{W_{\ell}}^{T}\right)=\left[\begin{array}{c}
\phi\left(\mathbf{x}_{1}^{T}\right) \\
\phi\left(\mathbf{x}_{2}^{T}\right) \\
\vdots \\
\phi\left(\mathbf{x}_{\ell}^{T}\right)
\end{array}\right]=\left[\begin{array}{c}
\mathbf{y}_{1}^{T} \\
\mathbf{y}_{2}^{T} \\
\vdots \\
r_{\ell} \times(J+1) \\
\mathbf{y}_{r_{\ell}}^{T}
\end{array}\right] .  \tag{16}\\
r_{\ell} \times J
\end{gather*} .
$$

$$
\begin{equation*}
\text { Let } \underset{r_{\ell} \times J}{\mathbf{Y}_{\ell}}=\tilde{\phi}\left(\mathbf{X}_{\ell}, W_{\ell}, D\right) \tag{17}
\end{equation*}
$$

Each row vector in $\mathbf{Y}_{\ell}$ is a vector of reals, all potentially from the multivariate normal distribution corresponding to the $\ell^{t h}$ pattern of zeros. Note that we cannot form a single block matrix, $\mathbf{Y}$, from the collection of $\mathbf{Y}_{\ell}$ because they can have different numbers of columns.

### 5.3 Means

The matrix $\mathbf{Y}_{\ell}$ contains rows of compositions with the same pattern of zeros. We refer to the $t^{t h}$ row vector of $\mathbf{Y}_{\ell}$ as $\mathbf{y}_{\ell t}^{T}$. We refer to the mean as the vector $\overline{\mathbf{y}}_{\ell}$, and define it as:

$$
\begin{equation*}
\underset{J \times 1}{\overline{\mathbf{y}}_{\ell}}=\frac{1}{r_{\ell}}\left(\mathbf{1}_{r_{\ell}}^{T} \mathbf{Y}_{\ell}\right)^{T} . \tag{18}
\end{equation*}
$$

Here we are using $\mathbf{1}_{r_{\ell}}$ to represent an $r_{\ell} \times 1$ column vector of ones. $\overline{\mathbf{y}}_{\ell}$ is a column vector. We define it this way because of how we intend to use it in quadratic forms from the multivariate normal density.

## 6 Simple Estimators

### 6.1 Mean

It is also possible to construct simpler estimators relying on properties of the normal distribution. For the location, if $\underset{n \times D}{\mathbf{X}}=\left[x_{t i}\right]$ is a collection of $n$ compositional data points with zeros, and the $D^{\text {th }}$ component always strictly positive, we can define a simple estimator of the mean, $\stackrel{*}{\boldsymbol{\mu}}=\left(\stackrel{*}{\mu_{1}}, \stackrel{*}{\mu}_{2}, \ldots, \stackrel{*}{\mu}_{d}\right)^{T}$. Let $n_{i}$ be the number of elements of the $i^{t h}$ column of $\mathbf{X}$ that are nonzero. For $i \in\{1,2, \ldots, d\}$, and $t \in\{1,2, \ldots, n\}$, define

$$
\begin{equation*}
\stackrel{*}{\mu_{i}}=\frac{1}{n_{i}} \sum_{\left\{t: x_{t i} \neq 0\right\}} \log \left(x_{t i} / x_{t D}\right) . \tag{19}
\end{equation*}
$$

By the assumption of normality of the logratios, the estimator $\stackrel{*}{\boldsymbol{\mu}}$ is unbiased. The maximum likelihood estimator, $\hat{\boldsymbol{\mu}}$, being unbiased, must have at least as small a variance, but this simple estimator, $\stackrel{*}{\boldsymbol{\mu}}$, may be more convenient in some cases, e.g., where the data set is large.

### 6.2 Variance

Here we show how to find estmators for variances and covariances using maximum likelihood estimators for normal random variables. For a single random composition, x, with
components $x_{1}, x_{2}, \ldots, x_{D}$, we substitute $\log \left(x_{i} / x_{D}\right)$ into the MLE for variances of normal random variables. We use $\stackrel{*_{\sigma}}{i i}$ for the estimator of the variances of the logratios $\log \left(x_{i} / x_{D}\right)$, for $i \in\{1,2, \ldots, d\}$, and $t \in\left\{1,2, \ldots, n_{i}\right\}$.

$$
\begin{equation*}
\stackrel{* 2}{\sigma}_{i i}=\frac{1}{n_{i}} \sum_{\left\{t: x_{t i} \neq 0\right\}}\left(\log \left(x_{t i} / x_{t D}\right)-\stackrel{*}{\mu}_{i}\right)^{2} \tag{20}
\end{equation*}
$$

If we want an unbiased estimator, we can divide by $\left(n_{i}-1\right)$ instead of $n_{i}$. As with means, the different $\stackrel{*}{\sigma}_{i i}$ are based on different numbers of observations, $n_{i}$.

### 6.3 Covariance

It only makes sense to talk about estimating the covariance of the variables $\log \left(x_{i} / x_{D}\right)$ and $\log \left(x_{j} / x_{D}\right)$ when both $x_{t i}$ and $x_{t j}$ are not 0 so we define $n_{i j}=\left\|\left\{t: x_{t i} \neq 0 \& x_{t j} \neq 0\right\}\right\|$. That is, $n_{i j}$ is the number of data points where both $x_{t i}$ and $x_{t j}$ are not 0 . As we did with variance, we can start with the canonical maximum likelihood formula for estimating covariance among normally distributed variables, and substitute in appropriate logratios.

$$
\begin{equation*}
\stackrel{*}{\sigma}_{i j}=\frac{1}{n_{i j}} \sum_{\left\{t: x_{t i} \neq 0 \& x_{t j} \neq 0\right\}}\left(\log \left(x_{t i} / x_{t D}\right)-\stackrel{*}{\mu}_{i}\right)\left(\log \left(x_{t j} / x_{t D}\right)-\stackrel{*}{\mu}_{j}\right) \tag{21}
\end{equation*}
$$

Note that $\stackrel{*}{\sigma}_{i j}$ is based on $n_{i j}$ observations, while $\stackrel{*}{\mu}_{i}$ and $\stackrel{*}{\mu}_{j}$ are based on $n_{i}$ and $n_{j}$ observations, respectively. The formula in Equation 21 is based on the maximum likelihood estimator for covariance of normal variables. For unbiased estimators we would divide by $\left(n_{i j}-1\right)$ instead of $n_{i j}$.

Our estimator for the $d \times d$ variance-covariance matrix is $\stackrel{*}{\boldsymbol{\Omega}}=\left[\stackrel{*}{\sigma}_{i j}\right]$.

## 7 Maximum Likelihood Estimators

For the case where there are no zeros, the location estimator described earlier is a maximum likelihood estimator, but in general the the estimator we found earlier is not an MLE. From now on we will call that estimator the simple estimator, to contrast it with the maximum likelihood estimator, which we derive next.

We start by finding the location MLE given $\boldsymbol{\Omega}$ for 3 -part compositions, show it is unbiased, and then show the relative efficiency of the simple estimator with to the MLE. Assume we have a set of logistic normal compositional data with $b$ different patterns of zeros:

$$
\begin{aligned}
& \mathbf{x}_{11}, \ldots, \mathbf{x}_{1 r_{1}} \stackrel{i . i . d .}{\sim} \mathscr{L}^{\left|W_{1}\right|}\left(\mathbf{B}_{W_{1}} \boldsymbol{\mu}, \mathbf{B}_{W_{1}} \boldsymbol{\Omega} \mathbf{B}_{W_{1}}^{T}\right) \\
& \mathbf{x}_{21}, \ldots, \mathbf{x}_{2 r_{2}} \stackrel{i . i . d .}{\sim} \mathscr{L}^{\left|W_{2}\right|}\left(\mathbf{B}_{W_{2}} \boldsymbol{\mu}, \mathbf{B}_{W_{2}} \boldsymbol{\Omega} \mathbf{B}_{W_{2}}^{T}\right)
\end{aligned}
$$

$$
\begin{align*}
& \vdots \tag{22}
\end{align*}
$$

In a block of data, as in 22 , we use $\mathbf{x}_{\ell t}$ to refer to the $t^{t h}$ composition of the compositions with $W_{\ell}$ for their pattern of zeros. We define $\mathbf{y}_{\ell t}=\tilde{\phi}\left(\mathbf{x}_{\ell t}, W_{\ell}, D\right)$, and to ease notation, we write in terms of $\mathbf{y}_{\ell t}$.

### 7.1 Likelihood

First we write the full likelihood and log likelihood for D-place compositions, and then restrict ourselves to 3 -place compositions. The full likelihood is:

$$
\begin{align*}
& L\left(\boldsymbol{\mu}, \boldsymbol{\Omega} \mid r_{1}, \ldots, r_{b}, \mathbf{y}_{11}, \ldots, \mathbf{y}_{b r_{b}}\right)=  \tag{23}\\
& \prod_{\ell=1}^{b} \prod_{t=1}^{r_{\ell}} \frac{P_{\ell}}{(2 \pi)^{\left|W_{\ell}\right| / 2}\left|\mathbf{B}_{W_{\ell}} \boldsymbol{\Omega} \mathbf{B}_{W_{\ell}}^{T}\right|^{1 / 2}} \exp \left[-\frac{1}{2}\left(\mathbf{y}_{\ell t}-\mathbf{B}_{W_{\ell}} \boldsymbol{\mu}\right)^{T}\left(\mathbf{B}_{W_{\ell}} \boldsymbol{\Omega} \mathbf{B}_{W_{\ell}}^{T}\right)^{-1}\left(\mathbf{y}_{\ell t}-\mathbf{B}_{W_{\ell}} \boldsymbol{\mu}\right)\right] .
\end{align*}
$$

The constant

$$
\begin{equation*}
\prod_{\ell=1}^{b} \prod_{t=1}^{r_{\ell}} \frac{P_{\ell}}{(2 \pi)^{\left|W_{\ell}\right| / 2}\left|\mathbf{B}_{W_{\ell}} \boldsymbol{\Omega} \mathbf{B}_{W_{\ell}}^{T}\right|^{1 / 2}} \tag{24}
\end{equation*}
$$

is independent of $\boldsymbol{\mu}$, so for purposes of maximizing the likelihood with respect to $\boldsymbol{\mu}$, we can treat it as a single constant, $C$.

$$
\begin{align*}
& L\left(\boldsymbol{\mu}, \boldsymbol{\Omega} \mid r_{1}, \ldots, r_{b}, \mathbf{y}_{11}, \ldots, \mathbf{y}_{b r_{b}}\right) \\
& =C \prod_{\ell=1}^{b} \prod_{t=1}^{r_{\ell}} \exp \left[-\frac{1}{2}\left(\mathbf{y}_{\ell t}-\mathbf{B}_{W_{\ell}} \boldsymbol{\mu}\right)^{T}\left(\mathbf{B}_{W_{\ell}} \boldsymbol{\Omega} \mathbf{B}_{W_{\ell}}^{T}\right)^{-1}\left(\mathbf{y}_{\ell t}-\mathbf{B}_{W_{\ell}} \boldsymbol{\mu}\right)\right]  \tag{25}\\
& =C \exp \left[-\frac{1}{2} \sum_{\ell=1}^{b} \sum_{t=1}^{r_{\ell}}\left(\mathbf{y}_{\ell t}-\mathbf{B}_{W_{\ell}} \boldsymbol{\mu}\right)^{T}\left(\mathbf{B}_{W_{\ell}} \boldsymbol{\Omega} \mathbf{B}_{W_{\ell}}^{T}\right)^{-1}\left(\mathbf{y}_{\ell t}-\mathbf{B}_{W_{\ell}} \boldsymbol{\mu}\right)\right] . \tag{26}
\end{align*}
$$

Taking the log gives:

$$
\begin{align*}
& \log L\left(\boldsymbol{\mu}, \boldsymbol{\Omega} \mid r_{1}, \ldots, r_{b}, \mathbf{y}_{11}, \ldots, \mathbf{y}_{b r_{b}}\right)  \tag{27}\\
& =\log C-\frac{1}{2} \sum_{\ell=1}^{b} \sum_{t=1}^{r_{\ell}}\left(\mathbf{y}_{\ell t}-\mathbf{B}_{W_{\ell}} \boldsymbol{\mu}\right)^{T}\left(\mathbf{B}_{W_{\ell}} \boldsymbol{\Omega} \mathbf{B}_{W_{\ell}}^{T}\right)^{-1}\left(\mathbf{y}_{\ell t}-\mathbf{B}_{W_{\ell}} \boldsymbol{\mu}\right) . \tag{28}
\end{align*}
$$

For the simple case of three-part compositional data with some zeros in component one, and some zeros in component two, the parent distribution for the transformed data is bivariate normal,

$$
\mathscr{N}(\boldsymbol{\mu}, \boldsymbol{\Omega}) \text { where } \boldsymbol{\mu}=\left[\begin{array}{l}
\mu_{1}  \tag{29}\\
\mu_{2}
\end{array}\right], \text { and } \boldsymbol{\Omega}=\left[\begin{array}{ll}
s_{11} & s_{12} \\
s_{12} & s_{22}
\end{array}\right]
$$

For the full bivariate normal distribution,

$$
A=\boldsymbol{\Omega}^{-1}=\frac{1}{s_{11} s_{22}-s_{12}^{2}}\left[\begin{array}{cc}
s_{22} & -s_{12}  \tag{30}\\
-s_{12} & s_{11}
\end{array}\right] .
$$

For the two univariate normal distributions, the inverses of the variances are: $\frac{1}{s_{11}}$ and $\frac{1}{s_{22}}$. In these formulas, $y_{1 j 1}$ is the $j^{\text {th }}$ data point among the univariate data from the first component. $y_{2 j 2}$ is the $j^{\text {th }}$ data point among the univariate data from the second component. $\mathbf{y}_{3 j}$ is a 2 -place vector with data from both components,

$$
\left[\begin{array}{c}
y_{1 j 1}  \tag{31}\\
\odot
\end{array}\right] \quad\left[\begin{array}{c}
\odot \\
y_{2 j 2}
\end{array}\right] \quad\left[\begin{array}{l}
y_{3 j 1} \\
y_{3 j 2}
\end{array}\right] .
$$

We also define their means.

$$
\frac{1}{r_{1}} \sum_{j=1}^{r_{1}} y_{1 j 1}=\bar{y}_{11} \quad \frac{1}{r_{2}} \sum_{j=1}^{r_{2}} y_{2 j 2}=\bar{y}_{22} \quad \frac{1}{r_{3}} \sum_{j=1}^{r_{3}} \mathbf{y}_{3 j}=\left[\begin{array}{l}
\bar{y}_{31}  \tag{32}\\
\bar{y}_{32}
\end{array}\right]
$$

### 7.1.1 Partial Derivatives

$$
\begin{align*}
& \frac{\partial \log L\left(\boldsymbol{\mu} \mid \mathbf{Y}, \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}\right)}{\partial \mu_{1}}= \\
& \frac{1}{s_{11}} r_{1}\left(\bar{y}_{11}-\mu_{1}\right)+\frac{s_{22}}{s_{11} s_{22}-s_{12}^{2}} r_{3}\left(\bar{y}_{31}-\mu_{1}\right)+\frac{-s_{12}}{s_{11} s_{22}-s_{12}^{2}} r_{3}\left(\bar{y}_{32}-\mu_{2}\right)  \tag{33}\\
& \frac{\partial \log L\left(\boldsymbol{\mu} \mid \mathbf{Y}, \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}\right)}{\partial \mu_{2}}= \\
& \frac{1}{s_{22}} r_{2}\left(\bar{y}_{22}-\mu_{2}\right)+\frac{s_{11}}{s_{11} s_{22}-s_{12}^{2}} r_{3}\left(\bar{y}_{32}-\mu_{2}\right)+\frac{-s_{12}}{s_{11} s_{22}-s_{12}^{2}} r_{3}\left(\bar{y}_{31}-\mu_{1}\right) \tag{34}
\end{align*}
$$

### 7.1.2 MLE for Location, Given $\Omega$

We set the partial derivatives equal to zero, replace $\mu$ with $\hat{\mu}$, and solve. The result is:

$$
\begin{align*}
& \hat{\mu}_{1} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}=\frac{\left(r_{1} \bar{y}_{11}+r_{3} \bar{y}_{31}\right)\left(r_{2}+r_{3}\right) s_{11} s_{12}-r_{1} \bar{y}_{11} r_{2} s_{12}^{2}+\left(\bar{y}_{22}-\bar{y}_{32}\right) r_{2} r_{3} s_{11} s_{12}}{\left(r_{1}+r_{3}\right)\left(r_{2}+r_{3}\right) s_{11} s_{22}-r_{1} r_{2} s_{12}^{2}}  \tag{35}\\
& \hat{\mu}_{2} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}=\frac{\left(r_{2} \bar{y}_{22}+r_{3} \bar{y}_{32}\right)\left(r_{1}+r_{3}\right) s_{11} s_{12}-r_{2} \bar{y}_{22} r_{1} s_{12}^{2}+\left(\bar{y}_{11}-\bar{y}_{31}\right) r_{1} r_{3} s_{12} s_{22}}{\left(r_{1}+r_{3}\right)\left(r_{2}+r_{3}\right) s_{11} s_{22}-r_{1} r_{2} s_{12}^{2}} \tag{36}
\end{align*}
$$

In the case where there are no univariate data from the second component, i.e, $r_{2}=0$, we have:

$$
\begin{equation*}
\left.\left(\hat{\mu}_{1} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}\right)\right|_{r_{2}=0}=\frac{r_{1} \bar{y}_{11}+r_{3} \bar{y}_{31}}{\left(r_{3}+r_{1}\right)}=\frac{1}{\left(r_{3}+r_{1}\right)}\left[\sum_{j=1}^{r_{3}} y_{3 j 1}+\sum_{j=1}^{r_{1}} y_{1 j 1}\right] . \tag{37}
\end{equation*}
$$

That shows that when we have $r_{2}=0$, the $\operatorname{MLE}\left(\hat{\mu}_{1} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}\right)$ is equal to our simple estimator for $\mu_{1}$. Similarly, when $r_{1}=0,\left(\hat{\mu}_{2} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}\right)$ is equal to our simple estimator for $\mu_{2}$. It also turns out that $\left.\left(\hat{\mu}_{1} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}\right)\right|_{r_{3}=0}=\bar{y}_{11}$, and when $r_{3}=0$, the simple estimator is also $\bar{y}_{11}$, so they are equal in that case as well.

### 7.2 Unbiasedness of Conditional MLE for 3-Part Composition

To show that $\hat{\boldsymbol{\mu}} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}$ is unbiased, we start by pointing out the expectations of the various means:

$$
\begin{array}{r}
E\left[\bar{y}_{11}\right]=E\left[\frac{1}{r_{1}} \sum_{j=1}^{r_{1}} y_{1 j 1}\right]=\frac{1}{r_{1}} \sum_{j=1}^{r_{1}} E\left[y_{1 j 1}\right]=\mu_{1} \\
E\left[\bar{y}_{22}\right]=E\left[\frac{1}{r_{2}} \sum_{j=1}^{r_{2}} y_{2 j 2}\right]=\frac{1}{r_{2}} \sum_{j=1}^{r_{2}} E\left[y_{2 j 2}\right]=\mu_{2} \\
E\left[\left[\begin{array}{l}
\bar{y}_{31} \\
\bar{y}_{32}
\end{array}\right]\right]=E\left[\frac{1}{r_{3}} \sum_{j=1}^{r_{3}} \mathbf{y}_{3 j}\right]=\frac{1}{r_{3}} \sum_{j=1}^{r_{3}} E\left[\mathbf{y}_{3 j}\right]=\left[\begin{array}{l}
\mu_{1} \\
\mu_{2}
\end{array}\right] \tag{40}
\end{array}
$$

When we take the expectation, the term with $\left(\bar{y}_{22}-\bar{y}_{32}\right)$ vanishes because $E\left[\bar{y}_{22}\right]=E\left[\bar{y}_{32}\right]$. That leaves only terms with $E\left[\bar{y}_{11}\right]=\mu_{1}$ and $E\left[\bar{y}_{31}\right]=\mu_{1}$, which we can factor.

$$
\begin{gather*}
\hat{\mu}_{1} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}=\frac{\left(r_{1} \bar{y}_{11}+r_{3} \bar{y}_{31}\right)\left(r_{2}+r_{3}\right) s_{11} s_{12}-r_{1} \bar{y}_{11} r_{2} s_{12}^{2}+\left(\bar{y}_{22}-\bar{y}_{32}\right) r_{2} r_{3} s_{11} s_{12}}{\left(r_{1}+r_{3}\right)\left(r_{2}+r_{3}\right) s_{11} s_{22}-r_{1} r_{2} s_{12}^{2}}  \tag{41}\\
E\left[\hat{\mu}_{1} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{2}\right]=\frac{\mu_{1}\left[\left(r_{1}+r_{3}\right)\left(r_{2}+r_{3}\right) s_{11} s_{12}-r_{1} r_{2} s_{12}^{2}\right]}{\left(r_{1}+r_{3}\right)\left(r_{2}+r_{3}\right) s_{11} s_{22}-r_{1} r_{2} s_{12}^{2}}=\mu_{1} \tag{42}
\end{gather*}
$$

This shows that $\hat{\mu}_{1}$ is unbiased. By symmetry we get that $\hat{\mu}_{2}$ is unbiased.

### 7.3 General Maximum Likelihood Estimators

For the general case of maximum likelihood estimators for higher dimensions than shown here, the log likelihood can be differentiated, and the score functions can be solved with a
computer algebra system. In addition, the Hessian can be checked to verify the solution is a maximum. We have done this for the case of 3-part compositions and don't anticipate any obstacles to extending the program to handle the general case of D-dimensions.

## 8 Variances of Location Estimators

Next we find variances of the two location estimators, the MLE, and the simple estimator. Both are unbiased. A question we need to answer is, what is the efficiency of the simple estimator relative to the MLE. We have been using $\hat{\boldsymbol{\mu}}$ for the MLE. We continue to use $\stackrel{*}{\boldsymbol{\mu}}$ for the simple estimator (of the location). In our discussion,

$$
\begin{equation*}
\operatorname{efficiency}\left(\mu_{1}, \hat{\mu}_{1}\right)=\frac{\operatorname{Var}\left(\hat{\mu}_{1}\right)}{\operatorname{Var}\left(\tilde{\mu}_{1}\right)} \tag{43}
\end{equation*}
$$

### 8.1 Variance of Location MLE, $\hat{\mu}$

Next we derive the variance of the location MLE, $\hat{\boldsymbol{\mu}} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}$. Recall that

$$
\begin{equation*}
\hat{\mu}_{1} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}=\frac{\left(r_{1} \bar{y}_{11}+r_{3} \bar{y}_{31}\right)\left(r_{2}+r_{3}\right) s_{11} s_{12}-r_{1} \bar{y}_{11} r_{2} s_{12}^{2}+\left(\bar{y}_{22}-\bar{y}_{32}\right) r_{2} r_{3} s_{11} s_{12}}{\left(r_{1}+r_{3}\right)\left(r_{2}+r_{3}\right) s_{11} s_{22}-r_{1} r_{2} s_{12}^{2}} . \tag{44}
\end{equation*}
$$

First we rewrite the expression so that each of the $\bar{y}$ terms stands alone.

$$
\begin{align*}
& \hat{\mu}_{1} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}= \\
& \frac{\left(r_{3}^{2}+r_{2} r_{3}\right) s_{11} s_{22} \bar{y}_{31}-r_{2} r_{3} s_{11} s_{12} \bar{y}_{32}+r_{2} r_{3} s_{11} s_{12} \bar{y}_{22}+\left(\left(r_{1} r_{3}+r_{1} r_{2}\right) s_{11} s_{22}-r_{1} r_{2} s_{12}^{2}\right) \bar{y}_{11}}{\left(r_{3}^{2}+\left(r_{2}+r_{1}\right) r_{3}+r_{1} r_{2}\right) s_{11} s_{22}-r_{1} r_{2} s_{12}^{2}} \tag{45}
\end{align*}
$$

To find the variance of $\hat{\mu}_{1} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}$, we need to replace $r_{1} \bar{y}_{11}$ with $\sum_{j=1}^{r_{1}} y_{1 j 1} ; r_{2} \bar{y}_{22}$ with $\sum_{j=1}^{r_{2}} y_{2 j 2}$; $r_{3} \bar{y}_{31}$ with $\sum_{j=1}^{r_{3}} y_{3 j 1}$; and $r_{3} \bar{y}_{32}$ with $\sum_{j=1}^{r_{3}} y_{3 j 2}$. We also make some other substitutions to simplify the algebra.

Let $k_{31}=\left(r_{3}+r_{2}\right) s_{11} s_{22}$.
Let $k_{32}=r_{2} s_{11} s_{12}$.
Let $k_{22}=r_{3} s_{11} s_{12}$.
Let $k_{11}=\left(r_{3}+r_{2}\right) s_{11} s_{22}-r_{2} s_{12}^{2}$.
Let $k_{\text {denom }}=\left(r_{3}^{2}+\left(r_{2}+r_{1}\right) r_{3}+r_{1} r_{2}\right) s_{11} s_{22}-r_{1} r_{2} s_{12}^{2}$.
With these in place, we get

$$
\hat{\mu}_{1} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}=\frac{1}{k_{\text {denom }}}\left(k_{31} \sum_{j=1}^{r_{3}} y_{3 j 1}-k_{32} \sum_{j=1}^{r_{3}} y_{3 j 2}+k_{22} \sum_{j=1}^{r_{2}} y_{2 j 2}+k_{11} \sum_{j=1}^{r_{1}} y_{1 j 1}\right)
$$

$$
\begin{equation*}
=\frac{1}{k_{\text {denom }}}\left(\sum_{j=1}^{r_{3}}\left(k_{31} y_{3 j 1}-k_{32} y_{3 j 2}\right)+k_{22} \sum_{j=1}^{r_{2}} y_{2 j 2}+k_{11} \sum_{j=1}^{r_{1}} y_{1 j 1}\right) . \tag{51}
\end{equation*}
$$

The $y_{2 j 2}$ are i.i.d. univariate normal; the $y_{1 j 1}$ are i.i.d. univariate normal; and the $\mathbf{y}_{3 j}$ are i.i.d bivariate normal, so the variance of the estimator is:
$\operatorname{Var}\left(\hat{\mu}_{1} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}\right)=$
$\left(\frac{1}{k_{\text {denom }}}\right)^{2}\left[\operatorname{Var}\left(\sum_{j=1}^{r_{3}}\left(k_{31} y_{3 j 1}-k_{32} y_{3 j 2}\right)\right)+\operatorname{Var}\left(k_{22} \sum_{j=1}^{r_{2}} y_{2 j 2}\right)+\operatorname{Var}\left(k_{11} \sum_{j=1}^{r_{1}} y_{1 j 1}\right)\right]$.
$\operatorname{Var}\left(y_{2 j 2}\right)=s_{22}$ and $\operatorname{Var}\left(y_{1 j 1}\right)=s_{11}$, so
$\operatorname{Var}\left(\hat{\mu}_{1} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}\right)=\left(\frac{1}{k_{\text {denom }}}\right)^{2}\left[\operatorname{Var}\left(\sum_{j=1}^{r_{3}}\left(k_{31} y_{3 j 1}-k_{32} y_{3 j 2}\right)\right)+k_{22}^{2} r_{2} s_{22}+k_{11}^{2} r_{1} s_{11}\right]$.

To find the variance of the remaining sum requires the facts that $\mathbf{y}_{3 j}$ are i.i.d., and that $\operatorname{Cov}\left(y_{3 j 1}, y_{3 j 2}\right)=s_{12}$.

$$
\begin{align*}
\operatorname{Var}\left(\sum_{j=1}^{r_{3}}\left(k_{31} y_{3 j 1}-k_{32} y_{3 j 2}\right)\right) & =\sum_{j=1}^{r_{3}} \operatorname{Var}\left(k_{31} y_{3 j 1}-k_{32} y_{3 j 2}\right) \\
& =\sum_{j=1}^{r_{3}}\left[\operatorname{Var}\left(k_{31} y_{3 j 1}\right)+\operatorname{Var}\left(k_{32} y_{3 j 2}\right)-2 k_{31} k_{32} \operatorname{Cov}\left(y_{3 j 1}, y_{3 j 2}\right)\right] \\
& =\sum_{j=1}^{r_{3}}\left[k_{31}^{2} s_{11}+k_{32}^{2} s_{22}-2 k_{31} k_{32} s_{12}\right] \\
& =r_{3}\left[k_{31}^{2} s_{11}+k_{32}^{2} s_{22}-2 k_{31} k_{32} s_{12}\right] \tag{54}
\end{align*}
$$

With that we can write the variance of the MLE, $\hat{\mu}_{1}$.

$$
\begin{align*}
& \operatorname{Var}\left(\hat{\mu}_{1} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}\right)= \\
& \left(\frac{1}{k_{\text {denom }}}\right)^{2}\left[r_{3}\left(k_{31}^{2} s_{11}+k_{32}^{2} s_{22}-2 k_{31} k_{32} s_{12}\right)+k_{22}^{2} r_{2} s_{22}+k_{11}^{2} r_{1} s_{11}\right] . \tag{55}
\end{align*}
$$

Substituting the values for the $k$ 's back in gives:
$\operatorname{Var}\left(\hat{\mu}_{1} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}\right)=$
$\frac{r_{3}\left(\left(r_{3}+r_{2}\right)^{2} s_{11}^{3} s_{22}^{2}+r_{2}^{2} s_{11}^{2} s_{12}^{2} s_{22}-2 r_{2}\left(r_{3}+r_{2}\right) s_{11}^{2} s_{12}^{2} s_{22}\right)+r_{2} r_{3}^{2} s_{11}^{2} s_{12}^{2} s_{22}+r_{1} s_{11}\left(\left(r_{3}+r_{2}\right) s_{11} s_{22}-r_{2} s_{12}^{2}\right)^{2}}{\left(\left(r_{3}^{2}+\left(r_{2}+r_{1}\right) r_{3}+r_{1} r_{2}\right) s_{11} s_{22}-r_{1} r_{2} s_{12}^{2}\right)^{2}}$

Symmetry also gives the variance of $\hat{\mu}_{2}$ given $\boldsymbol{\Omega}$.
$\operatorname{Var}\left(\hat{\mu}_{2} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}\right)=$
$\frac{r_{3}\left(\left(r_{3}+r_{1}\right)^{2} s_{22}^{3} s_{11}^{2}+r_{1}^{2} s_{22}^{2} s_{12}^{2} s_{11}-2 r_{1}\left(r_{3}+r_{1}\right) s_{22}^{2} s_{12}^{2} s_{11}\right)+r_{1} r_{3}^{2} s_{22}^{2} s_{12}^{2} s_{11}+r_{2} s_{22}\left(\left(r_{3}+r_{1}\right) s_{22} s_{11}-r_{1} s_{12}^{2}\right)^{2}}{\left(\left(r_{3}^{2}+\left(r_{2}+r_{1}\right) r_{3}+r_{1} r_{2}\right) s_{11} s_{22}-r_{1} r_{2} s_{12}^{2}\right)^{2}}$

Next we find the variance of the simple estimator for the location so we can compare the two.

### 8.2 Variance of Simple Location Estimator, $\underset{\mu}{\mu}$

Our simple estimator for the location is $\stackrel{*}{\boldsymbol{\mu}}=\left[\begin{array}{l}\mu_{1} \\ \stackrel{*}{\mu_{2}}\end{array}\right]$. Here we concern ourselves with $\operatorname{Var}\left(\stackrel{*}{\mu}_{1}\right)$ and then rely on symmetry to arrive at the variance of $\stackrel{*}{\mu}_{2}$.

$$
\begin{align*}
\stackrel{*}{\mu}_{1} & =\frac{1}{r_{1}+r_{3}}\left[\sum_{j=1}^{r_{1}} y_{1 j 1}+\sum_{j=1}^{r_{3}} y_{3 j 1}\right]  \tag{58}\\
\operatorname{Var}\left(\stackrel{*}{\mu}_{1}\right) & =\operatorname{Var}\left(\frac{1}{r_{1}+r_{3}}\left[\sum_{j=1}^{r_{1}} y_{1 j 1}+\sum_{j=1}^{r_{3}} y_{3 j 1}\right]\right) \\
& =\frac{1}{\left(r_{1}+r_{3}\right)^{2}} \operatorname{Var}\left(\sum_{j=1}^{r_{1}} y_{1 j 1}+\sum_{j=1}^{r_{3}} y_{3 j 1}\right) \\
& =\frac{1}{\left(r_{1}+r_{3}\right)^{2}}\left(r_{1} s_{11}+r_{3} s_{11}\right) \\
& =\frac{s_{11}}{r_{1}+r_{3}} \tag{59}
\end{align*}
$$

### 8.3 Relative Efficiency of Location Estimators

The first thing we show is that when the covariance element of $\boldsymbol{\Omega}$ is zero, i.e, $s_{12}=0$, then $\hat{\mu}=\stackrel{*}{\mu}$.
$\operatorname{Var}\left(\hat{\mu}_{1} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}\right)=$
$\frac{r_{3}\left(\left(r_{3}+r_{2}\right)^{2} s_{11}^{3} s_{22}^{2}+r_{2}^{2} s_{11}^{2} s_{12}^{2} s_{22}-2 r_{2}\left(r_{3}+r_{2}\right) s_{11}^{2} s_{12}^{2} s_{22}\right)+r_{2} r_{3}^{2} s_{11}^{2} s_{12}^{2} s_{22}+r_{1} s_{11}\left(\left(r_{3}+r_{2}\right) s_{11} s_{22}-r_{2} s_{12}^{2}\right)^{2}}{\left(\left(r_{3}^{2}+\left(r_{2}+r_{1}\right) r_{3}+r_{1} r_{2}\right) s_{11} s_{22}-r_{1} r_{2} s_{12}^{2}\right)^{2}}$

Evaluate at $s_{12}=0$.

$$
\begin{equation*}
\left.\operatorname{Var}\left(\hat{\mu}_{1} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}\right)\right|_{s_{12}=0}=\frac{r_{3}\left(\left(r_{3}+r_{2}\right)^{2} s_{11}^{3} s_{22}^{2}\right)+r_{1} s_{11}\left(\left(r_{3}+r_{2}\right) s_{11} s_{22}\right)^{2}}{\left(\left(r_{3}^{2}+\left(r_{2}+r_{1}\right) r_{3}+r_{1} r_{2}\right) s_{11} s_{22}\right)^{2}} \tag{62}
\end{equation*}
$$

Factor numerator and denominator.

$$
\begin{equation*}
\left.\operatorname{Var}\left(\hat{\mu}_{1} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}\right)\right|_{s_{12}=0}=\frac{\left(r_{3}+r_{2}\right)^{2} s_{11}^{3} s_{22}^{2}\left(r_{3}+r 1\right)}{\left(r_{3}+r_{1}\right)^{2}\left(r_{3}+r_{2}\right)^{2} s_{11}^{2} s_{22}^{2}}=\frac{s_{11}}{r_{3}+r_{1}}=\stackrel{*}{\mu_{1}} . \tag{63}
\end{equation*}
$$

Similarly,

$$
\begin{equation*}
\left.\operatorname{Var}\left(\hat{\mu}_{2} \mid \boldsymbol{\Omega}, r_{1}, r_{2}, r_{3}\right)\right|_{s_{12}=0}=\frac{s_{22}}{r_{3}+r_{2}}=\stackrel{*}{\mu_{2}} . \tag{64}
\end{equation*}
$$

We have already shown in Section 7.1.2 that when $r_{2}=0, \hat{\mu}_{1}=\stackrel{*}{\mu}_{1}$, and when $r_{1}=$ $0, \hat{\mu}_{2}=\stackrel{*}{\mu}_{2}$; and when $r_{3}=0, \hat{\mu}_{1}=\stackrel{*}{\mu_{1}}$, and $\hat{\mu}_{2}=\stackrel{*}{\mu_{2}}$. Next we need to compare the variance of $\stackrel{*}{\mu}$ with the variance of $\hat{\mu}$ in cases where the estimators are not obviously the same. To do this we wrote a program to calculate the ratios of the variances, $\operatorname{Var}\left(\hat{\mu}_{1}\right) / \operatorname{Var}\left(\mu_{1}\right)$ and $\operatorname{Var}\left(\hat{\mu}_{2}\right) / \operatorname{Var}\left(\mu_{2}\right)$ for each of the possible combinations of $r_{1}, r_{2}$, and $r_{3}$ such that $r_{1}+r_{2}+r 3=100$, for two different cases of $\boldsymbol{\Omega}$,

$$
\boldsymbol{\Omega}=\left[\begin{array}{cc}
1 & 0.2  \tag{65}\\
0.2 & 1
\end{array}\right] \text { and } \boldsymbol{\Omega}=\left[\begin{array}{cc}
1 & 0.8 \\
0.8 & 1
\end{array}\right] .
$$

In all three figures, we plot $\operatorname{Var}\left(\hat{\mu}_{2}\right) / \operatorname{Var}\left(\stackrel{*}{\mu}_{2}\right)$ versus $\operatorname{Var}\left(\hat{\mu}_{1}\right) / \operatorname{Var}\left(\stackrel{*}{\mu}_{1}\right)$. In Figure 1 we use a small covariance term, $s_{12}=0.2$. In Figure 2 we use a large covariance, $s_{12}=0.8$. In both figures, we shade by the size of $r_{1}$ relative to $r_{2}$.


Figure 1: Efficiency of $\stackrel{*}{\mu}$ Relative to $\hat{\mu}$ with Low Covariance (0.2)
Figure 1 shows the relationship between efficiency of $\stackrel{*}{\mu}_{1}$ and $\stackrel{*}{\mu}_{2}$ and the relative sizes of $r_{1}$ and $r_{2}$. In the worst case, when $r_{1} \gg r_{2}$, the efficiency of $\stackrel{*}{\mu}_{1}$ approaches 1 , and the efficiency of $\stackrel{*}{\mu}_{2}$ falls off toward 0.97 . A point to note here is that for a relatively small covariance, 0.2 , the simple estimator, $\stackrel{*}{\mu}$ has a variance almost as small as that of $\hat{\mu}$. We will save discussion of the bands or striations for Figure 3.


Figure 2: Efficiency of $\stackrel{*}{\mu}$ Relative to $\hat{\mu}$ with High Covariance (0.8)
Figure 2, which shows efficiency based on a covariance of 0.8 , has the same pattern as Figure 1, but with larger variances for $\stackrel{*}{\mu}$, smaller efficiency. Here the worst cases can have an efficiency of less than 0.5 for either component of $\stackrel{*}{\mu}$, though when the efficiency of $\stackrel{*}{\mu}_{1}$ is that small, the efficiency of $\stackrel{*}{\mu}_{2}$ is very near 1 .


Figure 3: Efficiency of $\stackrel{*}{\mu}$ Relative to $\hat{\mu}$ with High Covariance (0.8) and Relative to $r_{3}$
Figure 3 shows the same points, for a covariance of 0.8 , but shaded by the value of $r_{3}$. To help decipher it, we show a subset of the points in Figure 4.


Figure 4: Efficiency of $\stackrel{*}{\mu}$ Relative to $\hat{\mu}$ with High Covariance (0.8). $r_{3} \in\{1: 4,61: 64\}$.
Figure 4 shows a subset of the points, only the points where $r_{3} \in\{1,2,3,4,61,62,63,64\}$. When $r_{3}$ is very small there is a wide range of possibilities for $r_{1}$ and $r_{2}$. The four leftmost points in the upper left of Figure 4 are points where $r_{1}$ is 1 or $2 ; r_{2}$ is somewhere between 94 and 97 , and $r_{3}$ is 2,3 , or 4 . In these cases, the sample for estimating $\mu_{1}$ is very small, from 3 to 6 points, some from univariate data and some from the bivariate data. In that case, the MLE has a much smaller variance than the simple estimator. In that same case, there is a much larger sample from univariate data for estimating $\mu_{2}$, upwards of 90 points, plus a handful of points from the bivariate data. In that case, the difference between the variance of $\stackrel{*}{\mu}_{2}$ and $\hat{\mu}_{2}$ is very small.

Graphs with negative covariances, -0.2 , and -0.8 look the same as with positive covariances, and are omitted for the sake of brevity.

### 8.4 Summary of Relative Efficiency

Both the simple estimator for the location, $\stackrel{*}{\boldsymbol{\mu}}$, and the maximum likelihood estimator, $\hat{\boldsymbol{\mu}}$, are unbiased given $\boldsymbol{\Omega}$. The simple estimator's efficiency relative to the MLE tends to decrease as the covariance component of $\boldsymbol{\Omega}$ increases. We say "tends" because even with a covariance of 0.8 , there are cases where the efficiency of both components of $\stackrel{*}{\mu}$ relative to $\hat{\boldsymbol{\mu}}$ is very close to one.

When there are relatively few zeros, and they are balanced, $\stackrel{*}{\boldsymbol{\mu}}$ has a variance almost as small as $\hat{\boldsymbol{\mu}}$. The more zeros there are, or the more unbalanced their distribution is, the larger the variance of one or more components of the simple estimator.

## 9 Subcompositional Coherence

One of the key reasons for using the logistic normal approach is that, in the base case without zeros, it preserves subcompositional coherence. That is, parameter estimates based on a subcomposition will be consistent with estimates based on the whole composition (componentwise equality for shared components).

In the presence of zeros, do we maintain that property? It depends on which estimators are used. We have shown that in general when there are zeros, the MLE for the mean is not the same as the simple estimator for the mean. The maximum likelihood estimator does not preserve subcompositional coherence when we have zeros. The simple estimators, by construction, do preserve subcompositional coherence provided the same $D^{\text {th }}$ component is in both.

Thus for inference, there is a choice to be made between maintaining subcompositional coherence and maximizing likelihood. Issues of zeros and subcompositional coherence are discussed in Scealy and Welsh (2014).

## 10 Discussion

The goal has been to extend the additive logistic normal distribution to cope with essential zeros. We have done that by requiring that the final component of each composition be nonzero, and by projecting compositions with zeros onto smaller dimensional subspaces, thereby addressing the issues of division by zero, and the log of zero. We arrive at a mixture of logistic normals where each distribution has a mean and a covariance parameter which are projections from a common mean and covariance.

We construct two sets of estimators, simple estimators, $\stackrel{*}{\boldsymbol{\mu}} \stackrel{*}{\boldsymbol{\Omega}}$, and maximum likelihood estimators, $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Omega}}$. These are estimated using all of the compositions in the data, regardless of where the zeros occur, assuming only that the $D^{t h}$ component is always nonzero. The simple estimators preserve subcompositional coherence, while the maximum likelihood estimators do not.

In compositional data zeros are a common occurrence. We developed this logistic normal mixture model with the intention of making analysis of such data easier. For future work, we plan to extend existing compositional data methods for inference, graphing, clustering, etc., to work with this distribution.

## References

Aitchison, J. (1986). The statistical analysis of compositional data. Chapman \& Hall, Ltd.
Aitchison, J. and Kay, J. W. (2003). Possible solution of some essential zero problems in compositional data analysis. In Proceedings of CoDaWork 2003, The First Compositional Data Analysis Workshop. Universitat de Girona. Departament d'Informàtica i Matemàtica Aplicada. http://ima.udg.edu/Activitats/CoDaWork03/ paper_Aitchison_and_Kay.pdf.

Bacon-Shone, J. (2008). Discrete and continuous compositions.
Billheimer, D., Guttorp, P., and Fagan, W. F. (2001). Statistical interpretation of species composition. Journal of the American Statistical Association, 96(456), 1205-1214. http: //dx.doi.org/10.1198/016214501753381850.

Butler, A. and Glasbey, C. (2008). A latent gaussian model for compositional data with zeros. Journal of the Royal Statistical Society: Series C (Applied Statistics), 57(5), 505-520.

Daunis-i-Estadella, J. and Martín-Fernández, J., editors (2008). Proceedings of CODAWORK 2008, The 3rd Compositional Data Analysis Workshop. University of Girona. CD-ROM (ISBN: 978-84-8458-272-4).

Daunis-i-Estadella, J., Martín-Fernández, J., and Palarea-Albaladejo, J. (2008). Bayesian tools for count zeros in compositional data sets. In Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop, May 27-30, University of Girona, Girona (Spain), CD-ROM.

Fry, J. M., Fry, T. R., and Mclaren, K. R. (2000). Compositional data analysis and zeros in micro data. Applied Economics, 32(8), 953-959. http://www.tandfonline.com/doi/ pdf/10.1080/000368400322002.

Kent, J. T. (1982). The fisher-bingham distribution on the sphere. Journal of the Royal Statistical Society, B, 44, 71-80.

Leininger, T. J., Gelfand, A. E., Allen, J. M., and Silander Jr, J. A. (2013). Spatial regression modeling for compositional data with many zeros. Journal of Agricultural, Biological, and Environmental Statistics, 18(3), 314-334.

Martin-Fernandez, J. A., Palarea-Albaladejo, J., and Olea, R. A. (2011). Dealing With Zeros, pages 43-58. In Pawlowsky-Glahn and Buccianti (2011).

Palarea-Albaladejo, J. and Martin-Fernandez, J. (2008). A modified em alr-algorithm for replacing rounded zeros in compositional data sets. Computers $\xi^{G}$ Geosciences, $\mathbf{3 4}(8)$, 902-917.

Palarea-Albaladejo, J., Martín-Fernández, J. A., and Olea, R. A. (2014). A bootstrap estimation scheme for chemical compositional data with nondetects. Journal of Chemometrics, 28(7), 585-599.

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

Scealy, J. and Welsh, A. (2011). Regression for compositional data by using distributions defined on the hypersphere. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 73(3), 351-375.

Scealy, J. and Welsh, A. (2014). Colours and cocktails: Compositional data analysis 2013 lancaster lecture. Australian $\mathcal{E B}^{\text {New Zealand Journal of Statistics, 56(2), 145-169. }}$

Stewart, C. and Field, C. (2011). Managing the essential zeros in quantitative fatty acid signature analysis. Journal of Agricultural, Biological, and Environmental Statistics, 16(1), 45-69.

# Optimising Archaeologic Ceramics XRF Analyses 

J. Bergman ${ }^{1}$ and A. Lindahl ${ }^{2}$<br>${ }^{1}$ Dept. of Statistics, Lund University, Sweden; jakob.bergman@stat.lu.se<br>${ }^{2}$ Dept. of Geology, Lund University, Sweden


#### Abstract

We present the first results of an experiment which is aimed at ultimately producing recommendations for analysing archaeologic ceramics specimens using hand held XRF analysis devices. In the experiment we study the effects of different measurement durations, different number of measured points, and three different types of surface treatments (breakage, polished, grounded) when analysing ceramics specimens, while controlling for nine different types of clay and three different types of temper (no temper, sand, rock), in total almost 1000 analysed points. For each measurement, the proportions of 36 different elements and all other elements are estimated. In those cases with multiple measurements of a specimen, the compositional centre of the measurements is calculated. A complicating issue in the analysis is the large number of parts found to be below detection limit; 13 elements have more than $50 \%$ of the measurements below detection limit and for more than half of those (almost) all measurements are below detection limit. We try nine different strategies for imputing the values. Each estimated elemental composition is compared to a reference estimate using the simplicial distance. The log distances are finally analysed using analysis of variance with main and interaction effects. We find that the different surface treatments have the greatest effect on the distances: grounded specimens yield the most accurate estimates and polished surfaces the least. We also find a significant effect of increasing the number of measured points, but less effect of increasing the duration of the measurements.


Key words: Archaeologic XRF analyses, Archaeometric experiment, Ceramics analysis, Elemental composition analysis, Simplicial distance.

## 1 Introduction

X-ray fluorescence (XRF) analysis, using hand held devices, has gained increased popularity among archaeologist during the recent years, primarily because of its portability and relatively low cost. The analysis produces an estimate of the elemental composition of a specimen. There is, however, not any real knowledge or agreement of how the analyses should be done to obtain the best results. For how long time should a specimen be analysed? How many points on a specimen should be analysed? Which would be preferable, to analyse one point for four minutes or two different points for two minutes each?

To obtain a good measurement one also needs to consider the type of surface that is being analysed. When an archaeologic artefact is encountered, the surface has usually been exposed to various chemical and mechanical interactions with surrounding materials, changing the elemental composition of the surface. Hence the surface might not be representative of the rest of specimen. To overcome this one could ground the specimen to a fine powder which would mix all parts of the specimen and also remove any effect that large grains might have on the analysis. Another option would be to break off a small piece of the specimen to create a fresh breakage surface gaining access to the interior of the specimen. A third more controllable option would be to remove a part of the surface of the specimen by polishing it with a suitable tool. An important question is how the choice of treatment will affect the analysis. Is one alternative preferable to the others?

In an attempt to answer the questions above, we present some first results of an experiment in which we study the effects of number of points measured, measurement duration, and treatment of the surface. The design of the experiment is described in more detail in Section 2 and the results of the experiment are presented in Section 3.

## 2 Experimental design

Nine different, comercially available, clays were purchased. The clays are listed in Table 1. Each clay was partitioned into three parts and different types of temper was applied, i.e. different materials were added to the clays to control for shrinkage as is commonly done in pottery. Sand was added to the first partition, to the second partition crushed rock was added, and to the third partition no temper was added. From the in all 27 different clay partitions, small samples were produced resembling potsherds and fired in a modern kiln.

Table 1: The nine different clays used in the experiment.

| No. | Clay type | Description | Firing temp. |
| :--- | :--- | :--- | ---: |
| 1 | Earthenware | black | $970-1040^{\circ} \mathrm{C}$ |
| 2 | Earthenware | red, $25 \%$ grog 0.2 mm | up to $1220^{\circ} \mathrm{C}$ |
| 3 | Earthenware | pale red, mix of natural blue and red clay |  |
| 4 | Earthenware | white | $1020-1140^{\circ} \mathrm{C}$ |
| 5 | Earthenware | white, $25 \%$ grog $0-0.5 \mathrm{~mm}$ | $1000-1280^{\circ} \mathrm{C}$ |
| 6 | Stoneware | white | $1000-1300^{\circ} \mathrm{C}$ |
| 7 | Earthenware | red | $1000-1150^{\circ} \mathrm{C}$ |
| 8 | Stoneware | black, $40 \%$ grog 0- 0.5 mm | $1220-1260^{\circ} \mathrm{C}$ |
| 9 | Earthenware | red, all lime has been washed/removed |  |

From each of the 27 different types of potsherds three replicates was then prepared for analysis in one of three ways: one potsherd was broken to create a breakage surface commonly found in archaeologic ceramic samples, one potsherd was polished using a diamond polishing disc to give a "perfect", smooth surface, and one potsherd was grounded to a fine power to give a complete
mixture of the sample removing any differences between the surface and the interior of the potsherd. This produced in total 81 different samples. The elemental composition of each sample was then analysed at 1 and 5 points, during 60 and 380 seconds, yielding in theory 972 measurements. However, due to the human factor the number of points analysed were in a few cases four or six instead of five, yielding in total 971 measurement. The analysis was done using a portable XRF device providing measurements of 36 elements plus a "Balance" containing all other elements.

### 2.1 Measurements below detection limit

Looking at the measurements it was noted that a fairly large amount of measurements were below the detection limit (BDL). The number of BDL measurements for each element is given in Table 2. It should be noted that five elements (chlorine, cobalt, selenium, antimony, and bismuth) have more than $99 \%$ BDL measurements, and silver, cadmium, and tin have all more than $90 \%$ BDL. Furthermore, magnesium, nickel, copper, tungsten, and gold have more than $50 \%$ BDL measurements.

Table 2: The number of measurements below detection limit (BDL) for each element. In total the analysis comprises 971 measurements. Note that all measurements of selenium and bismuth were below detection limit, and all but one of antimony. Also chlorine, cobalt, silver, cadmium, and tin had more than $90 \%$ measurements below detection limit.

| Si | Ti | Al | Fe | Mn | Mg | Ca | K | P | S | V | Cr |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0 | 0 | 0 | 1 | 432 | 634 | 1 | 0 | 410 | 274 | 24 | 83 |
| Ni | Cu | Zn | Rb | Sr | Y | Zr | Nb | Ba | Pb | Th | Cl |
| 656 | 530 | 25 | 0 | 0 | 1 | 1 | 1 | 11 | 21 | 85 | 965 |
| Co | As | Se | Mo | Ag | Cd | Sn | Sb | W | Au | Bi | U |
| 968 | 315 | 971 | 221 | 931 | 922 | 918 | 970 | 658 | 854 | 971 | 422 |

It is of course problematic to analyse data with such a large amount of measurements BDL. At least two main strategies are conceivable: elements could be excluded or measurements could be imputed. As in all such cases, it becomes a question of retaining information but not altering the data too much. Imputing data is of course not a problem when only a limited number of measurements are imputed, but one can question the reasonableness of imputing almost all values. We have chosen to try different ways of excluding and imputing in order compare the effects of the different strategies. The procedure was done in two steps. First three data sets were created with all elements with more than $50 \%, 90 \%$, and $99 \%$ BDL, respectively, removed. In those cases where elements with observed measurements where excluded, the observed measurements were added to the balance. Secondly, three imputation schemes were implemented to each data set. A non-parametric imputation with 0.65 of the detection limit and a model-based lognormal with either fixed or random imputation values (Palarea-Albaladejo and Martín-Fernández 2013). All imputation was done using the functions multRepl and multLN in the R package $z$ Compositions (Palarea-Albaladejo and Martín-Fernández 2014). To provide some sort of comparisons of the effects of the imputation, the first two principal components for the nine data sets are plotted in Figure 1(A) and third and fourth are plotted in Figure 1(B). The plots in Figure 1(B) indicate that the major difference between the data sets might be between only retaining elements with less than $50 \% \mathrm{BDL}$ and keeping more elements. On can clearly see in Figure 1(A) that the random imputation adds more variation, as intended.

### 2.2 Assessing the accuracy of the measurements

In order to assess the accuarcy of the measurement a reference (or "truth") is needed. During the summer of 2013, we made an agreement with a colleague in Germany who had access to analytical



Figure 1: The (A) first and second and (B) third and fourth principal components from the compositional PCA analysis of the nine different data sets, with different colours for the different clays (see colourbar on the left) and different symbols for the different treatments: breakage surface (■), polished surface ( $\mathbf{\Delta}$ ), grounded ( 0 ). The results are rather which of course is to be expected. (The signs of some components have been reversed for some data sets for ease of comparison.)
equipment of greater accuracy to analyse the 27 different clay-temper combinations. To date we have not received the results, but hopefully the will arrive in the near future. However, we still needed a reference, so we decided to use the results we had. It was deemed that the grounded samples would provide the best estimates, and theory (and common sense?) suggested that longer duration would also provide better measurements. So, for every clay-temper combination the centre composition (Aitchison 1989)

$$
\mathcal{C}\left(g\left(x_{11}, \ldots, x_{n 1}\right), \ldots, g\left(x_{1 D}, \ldots, x_{n D}\right)\right)
$$

where $\mathcal{C}(\boldsymbol{x})=\left(x_{1}, \ldots, x_{D}\right) / \sum_{i=1}^{n} x_{i}$, i.e. the closure operation, and $g\left(x_{1}, \ldots, x_{n}\right)=\left(x_{1} \cdots x_{n}\right)^{1 / n}$, i.e. the geometric mean, of all the 380 seconds measurements was calculated. This was done separately for the nine different imputation schemes, thus obtaining nine different reference sets.

## 3 Analysis

For each combination of clay, temper, treatment, number of measured points, and measurement duration, we calculate the compositional centre of the measurements. Thus, for one measured point we keep that measurement and for five points we calculate the centre of the five points. This is repeated for all imputation schemes resulting in nine data sets of 324 compositional estimates. For each estimate we calculate the simplicial distance (Aitchison 1983, p. 64)

$$
d_{S}(\boldsymbol{x}, \boldsymbol{y})=\sqrt{\sum_{i=1}^{D}\left(\log \frac{x_{i}}{g(\boldsymbol{x})}-\log \frac{y_{i}}{g(\boldsymbol{y})}\right)}
$$

where $g(\cdot)$ denotes the geometric mean, from the corresponding reference composition, i.e. the composition of that combination of clay and temper, resulting in nine sets of 324 distances. The calculations are done using the R package compositions (van den Boogaart, Tolosana, and Bren 2014).

The logarithm of the distances are analysed using analysis of variance. (The logarithm of distances are used as the distances are only positive and are expected to have a skew distribution. The decision is further strengthened by the fact that Box-Cox transformations indicate an optimal $\lambda \approx 0.2$, i.e. fairly close to 0 .) We model the effect of different treatments, number of points and measurement duration including all two-way interactions, controlling for different clays and temper.

In Table 3 we present results of the analysis of variance for one of the imputation schemes, the fixed lognormal with less than $50 \%$ BDL. It can be noted that the clearly most significant factor is the treatment, i.e. if the measurement was done on a breakage surface, a polished surface, or on the grounded sample. The least significant factor is the duration of the measurement. The results are similar for the other imputation schemes. The only difference is that the duration becomes significant when the number of elements is increased.

To get an idea of how the distances differ for different factor levels, we calculate the change in estimated mean value for the various combinations of treatment, number of points and measurement duration compared to the baseline of one measured point for 60 seconds on a breakage surface. The values are presented in Table 4. The shortest distances are found when the samples are grounded and the longest for the polished surfaces, with the breakage surfaces in-between. An interesting observation is that, whereas the accuracy of the measurements are improved with increased measurement duration for the grounded sample, the accuracy deteriorates with increased measurement duration for polished surfaces. For increased number of measurements the, the accuracy is improved for 380 seconds duration for both breakage and polished surfaces but deteriorated for the shorter duration.

Figure 2 shows normal QQ plots of the residuals for each of the nine different analyses (imputation

Table 3: The results of the analysis of variance for lognormal fixed imputation retaining only elements with less than $50 \%$ BDL. The results for the other imputation schemes are similar. The treatment (breakage, polished or grounded) is the most significant factor and the duration of measurements the least significant. The main difference between the imputation schemes is that duration is significant when more elements are retained, but not when only elements with less than $50 \% \mathrm{BDL}$ are retained.

| Factor | Df | Sum Sq | Mean Sq | $F$ value | $p$-value |
| :--- | ---: | ---: | ---: | ---: | :--- |
| Clay | 8 | 4.478 | 0.560 | 0.9154 | 0.5038 |
| Temper | 2 | 1.616 | 0.808 | 1.3212 | 0.2683 |
| Treatment | 2 | 272.422 | 136.211 | 222.7604 | $0.0000^{* * *}$ |
| DURATION | 1 | 1.880 | 1.880 | 3.0746 | 0.0805 |
| Points | 1 | 10.266 | 10.266 | 16.7887 | $0.0001^{* * *}$ |
| Treatment:Duration | 2 | 19.024 | 9.512 | 15.5560 | $0.0000^{* * *}$ |
| Treatment:Points | 2 | 9.652 | 4.826 | 7.8924 | $0.0005^{* * *}$ |
| Duration:Points | 1 | 3.476 | 3.476 | 5.6848 | $0.0177^{*}$ |
| Residuals | 304 | 185.887 | 0.611 |  |  |

Significance codes: $0^{* * *} 0.001^{* *} 0.01^{*} 0.05$

Table 4: Changes in estimated mean value of the log distances for the various combinations of treatment, number of points and measurement duration. The changes are compared to the baseline of one measured point for 60 seconds on a breakage surface. The estimates come from the analysis of the lognormal fixed imputation data set retaining only elements with less than $50 \%$ BDL.

| Points | Duration |  | Treatment |  |
| :--- | :---: | :---: | :---: | :---: |
|  |  | Breakage | Polished | Grounded |
| 1 | 60 s | 0 | 0.26998 | -0.91252 |
|  | 380 s | 0.14817 | 0.86613 | -1.49240 |
| 5 | 60 s | 0.04198 | 0.41486 | -1.54592 |
|  | 380 s | -0.22417 | 0.59669 | -2.54012 |

schemes). The plots indicate that the residuals have a slightly skewed distribution possibly violating the normality assumption.

Apparently, the treatment has the greatest impact on the distances, and especially whether or not the sample was grounded. Since grounded samples were used to create the elemental reference compositions, we rerun the analyses without the grounded samples, i.e. with only breakage and polished surfaces. In all the nine data sets treatment and the interaction between treatment and duration are the only significant effects. As the results are similar for all data sets, we provide as an illustration in Table 5, the estimated changes in mean values of the log distances for the various combinations of treatment, number of points and measurement duration for the same data set as above, i.e. the lognormal fixed imputation with less than $50 \%$ BDL. We note that the breakage surface still provides shorter distances than the polished surface. An interesting observation is that for breakage surfaces, the mean distance decreases when either the number of points or the measurement duration is increased, but when both are increased not much is gained.

## 4 Conclusions and future research

We have investigated how the accuracy of the elemental composition analysis of clay specimens, using a hand held XRF analysis device, is affected by different types of surface treatments, different number of points measured, and different measurement durations. Our prior belief was that the


Figure 2: Normal QQ plots of the residuals for the nine different imputation schemes.
more points and the longer the duration, the better the accuracy. We also believed that a grounded specimen would produce the most accurate measurements. These suppositions are confirmed by the analyses.

We can conclude that a grounded specimen is the most important factor in obtaining an accurate measurement. We are actually a bit surprised by the large differences between the three surface treatments. The polished surface should provide an optimal surface for the XRF device, but turns out to yield the worst results. A possible explanation could be that the samples have been polluted by substances from the polishing disc, even though this seems unlikely. Another possible explanation could be that the polished surface prevents the analyst from avoiding the large grains of temper etc., which is easily done with a breakage surface. It is, however, a gratifying result that the breakage surface does so well. Grounding and, to a slightly lesser extent, polishing are both destructive treatments, which are often not an option for an archaeologist. One reason for the popularity of the handheld XRF device is that it can be used on artefacts in e.g. museums without damaging or even removing them. Grounding the specimen is thus an optimal but perhaps more theoretical strategy.

At least for breakage surfaces and grounded samples, the most accurate measurements are obtained when using both five points and a measurement duration of 380 seconds per point. (This is not the case for polished surfaces, which is rather puzzling.) However, from a practitioner's point of view, it should be noted that five measurements each during 380 seconds means that the total time of the analysis is more than 30 minutes, not counting the time setting up the equipment,

Table 5: Changes in estimated mean value of the log distances for the various combinations of treatment, number of points and measurement duration. The changes are compared to the baseline of one measured point for 60 seconds on a breakage surface. The estimates come from the analysis of the lognormal fixed imputation data set retaining only elements with less than $50 \%$ BDL.

| Points | Duration | Treatment |  |
| :--- | :---: | :---: | :---: |
|  |  | Breakage | Polished |
| 1 | 60 s | 0 | 0.26998 |
|  | 380 s | -0.12913 | 0.58883 |
| 5 | 60 s | -0.23531 | 0.13757 |
|  | 380 s | -0.22416 | 0.59670 |

preparing the specimen and moving the specimen between the analyses. Since time is limited (for most of us), the question becomes whether it is preferable to analyse a specimen at only one point for 380 seconds or at five points for 60 seconds each? (In each case the total time of analysis is about six minutes.) Our findings clearly show that measuring five points during 60 seconds yields more accurate estimates than one point during 380 seconds. It should be noted though, that for breakage and polished surfaces a single measurement of 60 seconds gives more accurate estimate when the grounded samples are included in the analysis.

This has been a first report from an ongoing experiment. It is of course highly unsatisfactory to use the same measurements that are analysed to estimate the reference values. We are therefore eagerly looking forward to getting new independently estimated reference values. In this paper an analysis of 971 measurements was presented. In total, the experiment to date consists of more than 1800 measurements and more than 139 hours of XRF analysis device running time. In order to obtain a balanced experiment, about half of the measurements were excluded, as not all combinations of the factors are currently measured. It remains thus to complete the measurement sequence. This will also allow us to identify any threshold values in number of points and measurement duration: How much is gained in accuracy when increasing the number of points from one to three compared to increasing the number of points from three to five? Is there an optimal combination of number of points and measurement duration? It also remains as future research to investigate why the breakage and polished surfaces yielded more accurate measurements when measured only once for 60 seconds, than compared to five measurement for 60 seconds and to one measurement for 380 seconds. Is there a reason for this, or is it some sort of artefact of the extremely strong treatment effect?

Our conclusions in this experiment so far are that an archaeologist intending to do an elemental analysis of a ceramic specimen using a hand held XRF analysis device should ground the specimen if possible, and if not possible find a fresh breakage surface, and analyse the specimen at five points for 60 seconds each or, time permitting, for 380 seconds each, and finally calculate the compositional centre of the measurements.

## References

Aitchison, J. (1983). Principal component analysis of compositional data. Biometrika 70(1), 57-65.
Aitchison, J. (1989). Measures of location of compositional data sets. Mathematical Geology 21(7), 787-790.
van den Boogaart, K. G., R. Tolosana, and M. Bren (2014). compositions: Compositional Data Analysis. R package version 1.40-1.
Palarea-Albaladejo, J. and J. A. Martín-Fernández (2013). Values below detection limit in com-
positional chemical data. Analytica Chimica Acta 764, 32-43.
Palarea-Albaladejo, J. and J. A. Martín-Fernández (2014). zCompositions: Imputation of Zeros and Nondetects in Compositional Data Sets. R package version 1.0.3.

# Advances in integrating isotopic data with compositional data analysis: applications for deep formation brine geochemistry 

M.S. Blondes ${ }^{1}$, M.A. Engle ${ }^{1,2}$, N.J. Geboy ${ }^{1}$<br>${ }^{1}$ U.S. Geological Survey, Eastern Energy Resources Science Center, Reston, VA, USA; mblondes@usgs.gov<br>${ }^{2}$ Dept. of Geological Sciences, University of Texas at El Paso, El Paso, TX, USA


#### Abstract

In the geosciences, radiogenic and stable isotope data are often used along with chemical concentrations to help determine sources of and processes affecting geochemical variation. Though some work combining isotopic data with concentration data has been done using compositional data analysis ( CoDa ), a number of questions remain. Tolosana-Delgado and others (2005) showed that concentrations (e.g. $\mathrm{SO}_{4}$ ) can be split into separate parts (e.g. ${ }^{34} \mathrm{SO}_{4},{ }^{32} \mathrm{SO}_{4}$ ) based on their isotopic ratios. They also showed that an isometric log-ratio (ilr) transform of a stable isotope ratio is proportional to values expressed in classical delta notation and therefore the delta notation can simply be scaled to compare isotopic data and log-ratio transformed elemental/ionic data simultaneously. Puig and others (2011) applied this result to discriminant analysis of groundwaters by separating centered log-ratio (clr) transformed compositional data from untransformed isotopic data and scaling both sets to an equal variance. We are expanding on this idea to show the utility of ilr transformations and clr-biplots for combining elemental data with not only stable isotope data, but also with radiogenic isotopes, which are generally not presented using delta notation.

In the present study, we use multiple data sets of deep formation brines and compare traditional mixing models to their CoDa counterparts to delineate fluid movement between reservoirs. Concentrations of individual isotopes are calculated using isotopic ratios and global mean isotopic abundances. One key result is that isotope parts (e.g. ${ }^{18} \mathrm{O},{ }^{17} \mathrm{O},{ }^{16} \mathrm{O},{ }^{2} \mathrm{H},{ }^{1} \mathrm{H},{ }^{87} \mathrm{Sr},{ }^{86} \mathrm{Sr}$ ) can simply be modelled by the major ion concentration $\left(\mathrm{H}_{2} \mathrm{O}, \mathrm{Sr}\right)$ in a clr-biplot as they are perfectly dependent. Another important result is that an ilr transformation of radiogenic isotope parts (e.g. ${ }^{86} \mathrm{Sr}$ and ${ }^{87} \mathrm{Sr}$ in ${ }^{87} \mathrm{Sr} /{ }^{86} \mathrm{Sr}$ ) can, like stable isotopes in delta notation, be treated as a linear function of the isotopic ratio, scaled only by a constant.


Key words: radiogenic isotopes, stable isotopes, produced waters, basin brines

## 1 Introduction

In the geosciences, variations in both radiogenic (i.e., isotope systems where one or more isotopes are created from the radioactive decay of a parent isotope) and stable (i.e., non-radiogenic) isotope ratios are often combined with variations in chemical concentrations to interpret geologic processes. Radiogenic isotope ratios (e.g. ${ }^{87} \mathrm{Sr} /{ }^{86} \mathrm{Sr}$ ) in minerals are a product of initial isotopic abundances, partitioning of elements during crystallization, and subsequent radioactive decay over time. Particular minerals and rock types often have a specific radiogenic isotopic signature that may be imparted onto waters that interact with them. Radiogenic isotopes therefore are often used to determine the origin of fluids from rocks with a known or assumed isotopic composition or to determine mixing relationships between multiple sources. Stable isotopes of a given element (e.g. ${ }^{16} \mathrm{O},{ }^{17} \mathrm{O},{ }^{18} \mathrm{O}$ or ${ }^{1} \mathrm{H},{ }^{2} \mathrm{H}$ ), on the other hand, partition from one another during numerous geologic and biologic processes. Variations in stable isotopes can record evidence of specific processes (e.g. evaporation) or mixing between fluids that previously underwent different geologic processes over time (e.g. mixing between isolated basin reservoir brines and fresh meteoric water).

Isotopic ratios are directly measured as proportions (e.g. ${ }^{18} \mathrm{O} /{ }^{16} \mathrm{O}$ ) using a mass spectrometer, whereas major cation (e.g. $\mathrm{Ca}, \mathrm{Na}$ ) and anion (e.g. $\mathrm{SO}_{4}, \mathrm{Cl}$ ) concentrations are measured as parts of a whole fluid. In traditional approaches, isotopic ratios are included together with concentrations during data analysis. However, if we want to combine isotopic ratios and concentration data in compositional data analysis ( CoDa ), particularly for multivariate analyses, it is important to have a comprehensive understanding of how to properly transform all relevant data into the Euclidean geometry of real space.

Previous studies (Tolosana-Delgado and others, 2005; Puig and others, 2011) have addressed the incorporation of stable isotopes into CoDa, but there has yet to be a comprehensive development of procedures for different CoDa techniques. Tolosana-Delgado and others (2005) showed that chemical concentrations (e.g. $\mathrm{SO}_{4}$ ) can be split into separate parts (e.g. ${ }^{34} \mathrm{SO}_{4},{ }^{32} \mathrm{SO}_{4}$ ) based on their isotopic proportions. They also showed that an ilr transform of a stable isotope ratio is proportional to values expressed in classical delta notation (expressed as per mil or parts per thousand relative to a standard) and therefore the delta notation can simply be scaled to compare isotopic data and major element concentrations simultaneously. Puig and others (2011) applied this result to discriminant analysis of groundwaters by separating clr transformed compositional data from raw isotopic data and scaled both to have equal variances. Here we show the utility of ilr transformations and clr-biplots for combining elemental data with not only stable isotope data, but also with radiogenic isotopes, which are generally not presented using delta notation.

This study aims to focus on three things:

1. Interpretation of isotopes using ilr transformed subcompositions
2. Interpretation of isotopes in clr-biplots
3. Discussion of scaling constituents with small log-ratio variances

To address these issues, we focus on isotopic and compositional data for produced waters, which are the waters co-generated with hydrocarbons during oil and gas development. Produced waters may include portions of water originally present in geologic formations (formation water) prior to oil and gas production, waters injected for well stimulation and hydraulic fracturing, and water condensing from the gas phase. In many cases the salinity of these fluids can exceed that of bulk modern seawater ( $\sim 35 \mathrm{~g} / \mathrm{L})$ by several times. Analyzing produced waters can help understand basin scale hydrogeology and the transport of injected fluids. Isotopic data are particularly useful in answering these types of questions because they can demonstrate whether fluids in different reservoirs have mixed, and therefore whether injected fluids have been transported from one reservoir to another. We use two comprehensive chemical and isotopic datasets of formation water to address CoDa of isotopes: 1) formation waters from potash mine shafts in Saskatchewan, Canada (Jensen and others, 2006) and 2) formation brines from a Permian salt dome in the North German Basin (Kloppman and others, 2001). We focus on two stable isotope systems (O and H) and one radiogenic isotope system $(\mathrm{Sr})$.

## 2 Treatment of isotopic data

Display of stable isotopic ratios on a manageable scale is often done using delta notation, in which the isotope ratio in question is normalized to a standard and multiplied by a factor of 1000. For example:

$$
\begin{equation*}
\delta^{18} \mathrm{O}=1000 * \frac{\left(\frac{{ }^{18} \mathrm{O}}{{ }^{16} \mathrm{O}}\right)_{\text {sample }}-\left(\frac{{ }^{18} \mathrm{O}}{{ }^{16} \mathrm{O}}\right)_{\text {standard }}}{\left(\frac{18}{16} \mathrm{O}\right)_{\text {standard }}} \tag{1}
\end{equation*}
$$

And

$$
\begin{equation*}
\delta D=\delta^{2} \mathrm{H}=1000 * \frac{\left(\frac{{ }^{2} \mathrm{H}}{{ }^{1} \mathrm{H}}\right)_{\text {sample }}-\left(\frac{{ }^{2} \mathrm{H}}{{ }^{1} \mathrm{H}}\right)_{\text {standard }}}{\left(\frac{{ }^{2} \mathrm{H}}{{ }^{1} \mathrm{H}}\right)_{\text {standard }}} \tag{2}
\end{equation*}
$$

In the case of high salinity samples, such as many produced waters, large differences exist between isotopic ratios of H and O depending on whether they were measured on a concentration (mass) basis or an activity basis. For correct conversion of the data into ilr or clr transformed results and for comparison with concentration data, use of O and H isotopic data on a concentration basis is critical. Results presented in an activity basis can be converted to a concentration basis using the methods of Sofer and Gat (1972; 1975). Strontium isotopic data are often presented as a simple isotopic ratio of the radiogenic to stable isotope ${ }^{87} \mathrm{Sr} /{ }^{86} \mathrm{Sr}$, not normalized to any standard. They are measured as a mass ratio and are therefore expressed on a concentration basis.

As Tolosana-Delgado and others (2005) point out, isotopic data in delta notation are proportional to logratios, but multiplied by a different scaling factor. Unlike $\mathrm{SO}_{4}$ concentration data that can be split into proportions by its isotopic ratio, O and H isotopes of water must be treated differently because the concentration of pure $\mathrm{H}_{2} \mathrm{O}$ in aqueous samples is not usually measured. Further, to do certain CoDa transformations it may be useful to treat each individual isotope of O and H as a separate concentration (e.g. ${ }^{16} \mathrm{O},{ }^{17} \mathrm{O},{ }^{18} \mathrm{O},{ }^{1} \mathrm{H},{ }^{2} \mathrm{H},{ }^{86} \mathrm{Sr},{ }^{87} \mathrm{Sr}$, etc.). To do this, in this study, we use the following procedure:

- All concentration data are presented in units of $\mathrm{mg} / \mathrm{kg}$ solution ( ppm ). The $\mathrm{H}_{2} \mathrm{O}$ concentrations were calculated as the difference between the samples' density and total dissolved solids (TDS) concentration, where TDS is equivalent to the sum of all dissolved ions in the fluid. For samples where density was not reported, it was estimated using its relationship with TDS known from brine samples from the Permian Basin of West Texas. Water concentrations are therefore perfectly dependent on the concentration of the other ions and inversely correlated with TDS.
- Based on stoichiometric relationships, $\mathrm{H}_{2} \mathrm{O}$ concentrations were converted to concentrations of O and H in water in the solution. For O and H isotopic data in delta notation, ${ }^{18} \mathrm{O} /{ }^{16} \mathrm{O}$ and ${ }^{2} \mathrm{H} /{ }^{1} \mathrm{H}$ ratios were determined from Equations 1 and 2, using the known composition of standard mean ocean water. Assuming the ${ }^{17} \mathrm{O} /{ }^{16} \mathrm{O}$ ratio is constant and that of the average global abundances, and assuming ${ }^{3} \mathrm{H}$ is negligible, the concentrations of ${ }^{16} \mathrm{O},{ }^{17} \mathrm{O},{ }^{18} \mathrm{O},{ }^{1} \mathrm{H}$, and ${ }^{2} \mathrm{H}$ were individually calculated. For Sr isotopes, the Sr concentration in $\mathrm{mg} / \mathrm{kg}$ solution is split into 4 isotopic concentrations $\left({ }^{84} \mathrm{Sr},{ }^{86} \mathrm{Sr},{ }^{87} \mathrm{Sr}\right.$, and ${ }^{88} \mathrm{Sr}$ ) using the average global abundances for ${ }^{84} \mathrm{Sr},{ }^{86} \mathrm{Sr}$, and ${ }^{88} \mathrm{Sr}$, and the measured ${ }^{87} \mathrm{Sr} /{ }^{86} \mathrm{Sr}$ ratio.


## 3 Interpretation of isotopic data

### 3.1 Interpretation of isotope data using ilr transformed subcompositions

Ilr transformation of subcompositions can be utilized to interpret brine geochemistry. Engle and Rowan (2013) showed that not only can ilr transformations of the $\mathrm{Na}-\mathrm{Cl}-\mathrm{Br}$ system reveal halite dissolution and seawater evaporation trends common in certain environments that are apparent in traditional ratio plots,
but they also more clearly show secondary processes such as albitization or incorporation of clays. Engle and Blondes (2014) used ilr transformations of produced waters in the Permian Basin, USA, to link geochemical variation to thermodynamic mineral equilibration models. The incorporation of isotopes into such studies would be an important advancement for the study of basinal brine geochemistry.

Figure 1A shows a common plot of hydrogen and oxygen isotopes in water for brines from the North German Basin (Kloppman and others, 2001) and from potash mine shafts in Canada (Jensen and others, 2006). The global meteoric water line (GMWL) represents the isotopic composition of terrestrial surface and near surface waters (Craig, 1961). The position of a given data point along that line is a function of many variables, dominantly latitude, elevation, and temperature. Evaporative loss results in "heavier" $\delta^{18} \mathrm{O}$ values that plot to the right of the GMWL. Deep formation brines that previously experienced evaporation but have not interacted with surface waters tend to plot here.

The Permian North German Basin brines (Kloppman and others, 2001) plot directly on the GMWL, suggesting they are of meteoric origin. The formation waters from the potash mine shafts in Saskatchewan (Jensen and others, 2006) have a wide range in isotopic composition that is nearly as broad as all known terrestrial water compositions (Rozanski and others, 1993). This makes it particularly useful to detect differences between the traditional plot and a similar one created using the ilr transform of a four-part subcomposition, with each part being the individual isotopic concentrations. At first glance, the plots look identical except for the values on the axes, suggesting that the difference is simply one of scale. However, it is apparent from a close look at Figure 1B that the GMWL shows some curvature at the top right of the plot. Though these differences do exist, given that these data span nearly the entire range in observed values for the isotopic composition of natural waters it is small enough to be negligible and the ilr transformation does not appear to offer any benefits in this type of plot.

We can also examine radiogenic $\left.{ }^{87} \mathrm{Sr}\right)^{86} \mathrm{Sr}$ isotopes in a similar way. Figure 2 A again shows the Kloppman and others (2001) data set, color-coded by water type where fresh is $<1 \mathrm{~g} / \mathrm{L}$ TDS, brackish is $1-10 \mathrm{~g} / \mathrm{L}$ TDS, saline is $10-100 \mathrm{~g} / \mathrm{L}$ TDS, and brine is $>100 \mathrm{~g} / \mathrm{L}$ TDS. This plot of ${ }^{87} \mathrm{Sr}{ }^{86} \mathrm{Sr}$ vs. $1 / \mathrm{Sr}$ is a standard way to show mixing between Sr isotopic compositions, as pure mixtures of two end-members will fall along straight lines in this type of plot. The black line in this figure is a hypothetical linear mixture between the most fresh and the most saline (brine) sample in the data set. One can quickly infer that these samples do not derive their chemical and isotopic variation from simple mixing of these two endmembers. Using a two-part ilr transformation for the Sr isotopes ${ }^{87} \mathrm{Sr}$ and ${ }^{86} \mathrm{Sr}$ (Fig. 2B), the same scaling relationship observed in Figure 1B is apparent, whereby the log-ratio of a small variation is simply that small variation scaled. Figure 2C shows a three-part ilr transformation with ${ }^{87} \mathrm{Sr},{ }^{86} \mathrm{Sr}$, and $\mathrm{H}_{2} \mathrm{O}$. Similar interpretations of mixing can be done by comparing the data to the models in real space, yet the mixing relationship is no longer linear.


Figure 1: H and O isotopes of water, presented $(\mathrm{A})$ in standard delta notation and $(\mathrm{B})$ as a four-part ilr transformation. GMWL is the global meteoric water line. Note the slight curvature of the GMWL in figure 1B.


Figure 2: Sr isotope plots of brines. (A) Traditional presentation of strontium isotopes in which mixing is linear. (B) Two-part ilr transformation for the y-axis using ${ }^{87} \mathrm{Sr}$ and ${ }^{86} \mathrm{Sr}$. (C) Three-part ilr transformation using ${ }^{87} \mathrm{Sr},{ }^{86} \mathrm{Sr}$, and $\mathrm{H}_{2} \mathrm{O}$.

### 3.2 Interpretation of isotopes using clr-biplots

Principal component analysis of clr-biplots are useful for showing multivariate compositional data on a single plot (Aitchison and Greenacre, 2002). When derived from the covariance matrix of clr-transformed variables, the length of the links between the ray end points approximate the relative logratio variance between those two variables. Figure 3A shows covariance clr-biplot of only TDS and all of the O and H isotopes of water. The $\mathrm{H}_{2} \mathrm{O}$ concentration is also shown for illustrative purposes to compare the individual isotope rays to the $\mathrm{H}_{2} \mathrm{O}$ ray, but it is a redundant variable since $\left[\mathrm{H}_{2} \mathrm{O}\right]=\left[{ }^{18} \mathrm{O}+{ }^{17} \mathrm{O}+{ }^{16} \mathrm{O}+{ }^{2} \mathrm{H}+{ }^{1} \mathrm{H}\right]$. All isotopes of O and H in water lie on the same ray as $\mathrm{H}_{2} \mathrm{O}$, which is unsurprising as their concentrations are completely dependent on one another and sum to $\mathrm{H}_{2} \mathrm{O}$. The scores are more clearly seen in the "form" version of the biplot (Fig. 3B), in which the relationship between the scores and the coordinate axes are more visible. Here we see that the 1 st coordinate axis contains the entirety of the variability and again there is a clear salinity trend. Though the rays and the links between them have less meaning in the form biplot, small deviations are visible between the numerators in the isotopic ratios $\left({ }^{18} \mathrm{O},{ }^{2} \mathrm{H}\right)$ and the ray that represents $\mathrm{H}_{2} \mathrm{O},{ }^{16} \mathrm{O},{ }^{17} \mathrm{O}$, and ${ }^{1} \mathrm{H}$. The longer ray between ${ }^{1} \mathrm{H}$ and ${ }^{2} \mathrm{H}$ relative to the ray between ${ }^{18} \mathrm{O}$ and ${ }^{16} \mathrm{O}$ on the biplot confirms theory and observation that the larger mass differences between isotopes in the hydrogen system ( ${ }^{2} \mathrm{H}$ is has twice the mass of $\left.{ }^{1} \mathrm{H}\right)$ produces more isotopic fractionation than the oxygen system ( ${ }^{18} \mathrm{O}$ only has $12.5 \%$ more mass than ${ }^{16} \mathrm{O}$ ). Nearly $100 \%$ of the variance is found in the link between TDS and $\mathrm{H}_{2} \mathrm{O}$ and the scores project along this link as one would expect according to salinity. Comparison of the samples scores with the link between the water isotopes confirms that with a couple of exceptions, the fresh and brackish water samples tend to be isotopically heavy, and as salinity increases the samples generally get isotopically lighter. This type of plot thus appears to provide insights into connections between water salinity and O and H isotopic compositions.

## 4 Discussion and Conclusions

Interpreting variations in isotopic ratios is a valuable tool in the geosciences. A great number of natural processes can generate measurable differences in isotopic composition that are supported by models and experiments. However, though isotopes in a set of data may vary greatly relative to their mean, their variation is minor relative to the major components in the system and become negligible in multivariate clr approaches. The small relative variance is partially the reason that conversion of isotopic to ratio to ilr balances yields little change. As such, ilr transformations of subcompositions of sets of isotopes allow isotopic variation to be visible, but their interpretations and benefits when compared to simple logratios are unclear. However, as Tolosana-Delgado and others (2005) and Puig and others (2011) have shown for stable isotopes and we have shown here for radiogenic isotopes, familiar isotopic compositions either in delta notation or as simple ratios can be plotted against ilr transformed data without any loss of information; there is only a difference in scale.


Figure 3: clr-biplots of Kloppman and others (2001) data. Only TDS, H2O and the O and H isotopes are plotted. (A) Covariance biplot. (B) Form biplot.

The small log-ratio variations of isotopic data are swamped by major element compositions in clr-biplots. For multivariate approaches, either the data need to be scaled to a common variance (Puig and others, 2011) or another approach must be taken in data pre-processing or interpretation. In the case of O and H in brines, the isotopic system is largely controlled by the TDS/ $\mathrm{H}_{2} \mathrm{O}$ ratio, which suggests that the clrbiplot is a useful tool for examining differences in isotopic compositions of water as a function of salinity, an important characteristic of brines.

## References

Aitchison, J., Greenacre, M. (2002) Biplots of compositional data. Appl. Statist. 51:375-392.
Craig, H. (1961). Isotopic variations in meteoric waters. Science 133:1702-1703.
Engle, M.A., Blondes, M.S. (2014) Linking compositional data analysis with thermodynamic geochemical modeling: Oilfield brines from the Permian Basin, USA. Journal of Geochemical Exploration 141:61-70.

Engle, M.A., Rowan, E.L. (2013) Interpretation of $\mathrm{Na}-\mathrm{Cl}-\mathrm{Br}$ systematics in sedimentary basin brines: Comparison of concentration, element ratio, and isometric log-ratio approaches. Mathematical Geosciences 45:87-101.

Jensen, G.K.S., Rostron, B.J., Duke, M.J.M., Holmden, C. (2006) Chemical profiles of formation waters from potash mine shafts, Saskatchewan; in Summary of Investigations 2006, Volume 1, Saskatchewan Geological Survey, Sask. Industry Resources, Misc. Rep. 2006-4.1, CD-ROM, Paper A-7, 8p.

Kloppman, W., Négrel, Ph., Casanova, J., Klinge, H., Schelkes, K., Guerrot, C. (2001) Halite dissolution derived brines in the vicinity of a Permian salt dome ( N German Basin). Evidence from boron, strontium, oxygen, and hydrogen isotopes. Geochimica et Cosmochimica Acta 65:4087-4101.

Puig, R., Tolosana-Delgado, R., Otero, N., Folch, A. (2011) Combining isotopic and compositional data: a discrimination of regions prone to nitrate pollution. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), Compositional Data Analysis: Theory and Applications. pp. 302-317. John Wiley \& Sons, Ltd.

Rozanski, K., Araguás-Araguás, L., Gonfiantini, R. (1993) Isotopic patterns in modern global precipitation. In P.K. Swart, K.C. Lohmann, J. McKenzie, and S. Savin (Eds.), Climate change in continental isotopic records, pp. 1-36. Washington, D.C.: American Geophysical Union.

Sofer Z., Gat, J.R. (1972) Activities and concentrations of oxygen-18 in concentrated aqueous salt solutions: analytical and geophysical implications. Earth and Planetary Science Letters 15:232-238.

Sofer Z., Gat, J.R. (1975) The isotope composition of evaporating brines: effect of the isotopic activity ratio in saline solutions. Earth and Planetary Science Letters 26:179-186.

Tolosana-Delgado, R., Otero, N., and Soler, A. (2005) A compositional approach to stable isotope data analysis. 2nd Compositional Data Analysis Workshop CoDaWork '05, 11p.

# Linear models for the conditional distribution 

K. Gerald van den Boogaart ${ }^{1,2}$, Martin Lehmann ${ }^{2}$ R. Tolosana Delgado ${ }^{1}$<br>${ }^{1}$ Helmholtz Zentrum Dresden Rossendorf, Helmholtz Institute Freiber for Resource Technology, Germany, boogaart@hzdr.de<br>${ }^{2}$ TU Bergakademie Freiberg, Institute for Stochastics, Germany


#### Abstract

Our aim is to introduce a regression model for the conditional probablity distributions, of observable random variables as a function of regressors. The approach linear models in Bayes spaces. Bayes spaces are a generalisation of the Aitchison Simplex, which can be seen as the space of all distributions with support $1, \ldots, n$ to probablity distributions on arbitrary measure spaces. Compositional linear models on the Aitchison simplex turn out to be equivalent to multivariate logistic regression if written in alr representation. Then, likelihood theory for generalized linear models is used to estimate the regression coefficients based on observed realisation of the random variable instead of observed compositions as in compositional regression. This idea is generalized towards Bayes spaces, generating a technique allowing to define and estimate models for the conditional distribution of an observed quantity as a Bayes-space-linear function of regressors.


Key words: Multinumial Logistic Regression, Bayes Spaces, Compositional Regression

## 1 Introduction

Bayes Hilbert spaces generalize the Aitchison Simplex towards a vector space of probability distributions. Using the Aichison vector space structure of the simplex it was possible to generalize multivariate linear regression, analysis of variance and linear models to compositional response variables with compositional coefficients (van den Boogaart and Tolosana 2013, page 95-175). Bayes spaces allow to generalize this further to distributional response variables (van den Boogaart, Egozcue, and Pawlowsky-Glahn 2014) using distributional coefficients. We take this one step further towards linear modelling of the distribution of the response variable by a linear model in Bayes spaces. This is done by identifying logistic regression and multinomial logistic regression as linear models in the Aitchison simplex and then generalizing them towards Bayes spaces, i.e. a model which regression coefficients are objects from a Bayes Space (Lehmann, 2014). The result allows to mathematically describe the whole conditional distribution by a linear model in Bayes spaces. This approach has been described mathematically in (Lehmann 2014) and sucessfully used in (van den Boogaart, K. G. et al. 2014) for a novel geostatistical simulation procedure. This contribution discusses practical considerations and estimation for using it in general purpose regression. First we need to explain the model itself.

### 1.1 Multinomial logistic regression

Logistic regression formulates the conditional probability that an event takes place by a linear model

$$
\operatorname{logit}\left(p_{1}\right)=\sum_{j=1}^{p} b_{j} f_{j}(X)
$$

for the logit transform

$$
\operatorname{logit}(p)=\ln \frac{p}{1-p}
$$

where $p_{1}=P(Y=1 \mid X)$ is the conditional sucess probablity of a conditional binary response $Y$ given contitions $X$ and $f_{j}$ being known functions, such as $f_{1}(x) \equiv 1$ or $f_{2}(x)=x_{1}$, as in any general linear model. Typically an intercept $f_{1}(x) \equiv 1$ is assumend to be present in the model. Considering the probablity distribution for the binary response variable $Y$ to as a composition $\mathbf{p}=\left(p_{1}, 1-p_{1}\right)$ allows to replace the logit with $\operatorname{alr}(\mathbf{p})=\ln \frac{p_{1}}{p_{2}}=\ln \frac{p_{1}}{1-p_{1}}=\operatorname{logit}\left(p_{1}\right)$ :

$$
\operatorname{logit}\left(p_{1}\right)=\operatorname{alr}(\mathbf{p})=\sum_{j=1}^{p} f_{j}(X) b_{j}
$$

Multinomial logistic regression now replaces the logit by a log ratio against a special component for probablity:

$$
\begin{aligned}
\ln \frac{p_{1}}{p_{d}} & =\operatorname{alr}(\mathbf{p})_{1}=\sum_{j=1}^{p} f_{j}(X) b_{j 1} \\
& \vdots \\
\ln \frac{p_{d-1}}{p_{d}} & =\operatorname{alr}(\mathbf{p})_{d-1}=\sum_{j=1}^{p} f_{j}(X) b_{j d-1}
\end{aligned}
$$

for $k=1, \ldots, d-1$, which are the components of an alr transform. We can thus read this as a equation in alr transforms:

$$
\operatorname{alr}(\mathbf{p})=\sum_{j=1}^{p} f_{j}(X) \operatorname{alr}\left(\beta_{j}\right)
$$

where $\beta_{j}:=\operatorname{alr}^{-1}\left(\left(b_{j k}\right)_{k}\right)$. Applying $\operatorname{alr}^{-1}$ on both sides we get a perfect compositional linear model

$$
\mathbf{p}={ }_{A} \bigoplus_{j=1}^{p} f_{j}(X) \odot \beta_{j}
$$

where the $\beta_{j}$ are compositional parameters of a compositional linear model. These parameters can be estimated just like usual linear regression through a maximum likelihood estimator. These are computed in a usual way for generalized linear models by the Fisher scoring method.

While the model equation looks perfectly like the general form of a compositional regression model, the data observed from it is different. In a compositional regression model we observe a composition as dependent variable, i.e. often a normal distribution on the simplex with some residual covariance matrix $\mathbf{Y} \sim \mathcal{N}_{\mathcal{S}}^{D}\left(\operatorname{alr}(\mathbf{p}), \mathbf{S}_{\epsilon}\right)$. On the other hand, here we observe the realisation of a random variable distributed according to the resulting composition, i.e often a multinomial variate $\mathbf{Y} \sim$ $\mathcal{M} u(\mathbf{p}, n)$.

### 1.2 Bayes Hilbert Spaces

We would now like to generalize this concept towards a general regression for distributions, where the dependent variable might come from an arbitrary sample space, discrete, continuous, or even compositional. This is done by generalizing the approach from the Aitchison simplex to a Bayes space.

The Bayes space $B^{2}\left(P_{0}\right)$ is a space of all $\sigma$-finite measures mutually absolutely continuous to the measure $P_{0}$ having an log-square-integrable $P_{0}$-density. (i.e. having a strictly positive RadonNýkodim derivative with respect to $P_{0}$, which logarithm is square integrable), identified via a Bayes-equivalence relationship:

$$
\mu={ }_{B\left(P_{0}\right)} \nu: \Leftrightarrow \exists c \in R^{+}: \forall A: \mu(A)=c \nu(A)
$$

The Bayes-equivalence corresponds to the scale equivalence principle of Aitchison (1997) for compositions. As for compositions, we have two operations perturbation $\oplus$ and power transform $\odot$ making the Bayes space a vector space. These operations are defined as

$$
(\mu \oplus \nu):=_{B\left(P_{0}\right)} \int_{A} \frac{d \mu(x)}{d P_{0}(x)} \frac{d \nu(x)}{d P_{0}(x)} d P_{0}(x)
$$

and

$$
(\alpha \odot \mu):=_{B\left(P_{0}\right)} \int_{A}\left(\frac{d \mu(x)}{d P_{0}(x)}\right)^{\alpha} d P_{0}(x)
$$

$P_{0}$ is the neutral element of this vector space. A clr transform is definded as a mapping to the functions with $P_{0}$ integral 0:

$$
\operatorname{clr}(\mu)(x)=\ln \frac{d \mu(x)}{d P(x)}-\int \ln \frac{d \mu(x)}{d P(x)} d P_{0}(x)
$$

The inner product of the Bayes Hilbert Space is defined by the $L^{2}\left(P_{0}\right)$ inner product of the clrtransforms.

An Aitchison simplex is a special case of a Bayes-Hilbert-Space, where $P_{0}$ is the counting measure on the set of components. In general, Bayes Hilbert are infinite dimensional and do contain improper priors along with distribution. However, they are finite dimensional and only contain distributions if $P_{0}$ has a finite support, as in the case of the simplex. See (van den Boogaart et al. 2014) for details on Bayes Hilbert spaces.

## 2 Bayes space regression model

### 2.1 The model equation

We can now formally reinterpret the model equation of a compositional regression model in Bayes spaces:

$$
\begin{equation*}
P^{Y \mid X}=B\left(P_{0}\right) \bigoplus_{j=1}^{p} f_{j}(X) \odot \beta_{j} \tag{1}
\end{equation*}
$$

Here $P^{Y \mid X}$ is a conditional distribution and the regression parameters $\beta_{j} \in B^{2}\left(P_{0}\right)$ are now arbitrary elements of the Bayes space, i.e. measures. For a fixed choice of parameters this model would allow to compute a conditional distribution for every value of $X$.

### 2.2 Practical Challenges

Using this model as a statistical model introduces some practical challenges. First, the allowable space for $\beta_{j}$ is now an infinite dimensional space, while the observations are only a finite set of values from a finite dimensional space. Multiple different choices for the $\beta_{j}$ will thus fit equally and arbitrarily well. The model is too flexible and overparametrized. Our proposal is to restrict the model to a finite dimensional space, which is rich enough to model various distributions, but at the same time has a reasonably low number of parameters. Our proposal is based on the usual estimation process of functions in numerical computations by replacing the full space of functions by a smaller subspace of approximating functions. I.e. we find a basis of measures with:

$$
\beta_{j}=\bigoplus_{l=1}^{N} c_{j l} \odot \alpha_{j}
$$

and get regression model with real parameters:

$$
P^{Y \mid X}={ }_{B\left(P_{0}\right)} \bigoplus_{j=1}^{p} f_{j}(X) c_{j l} \odot \alpha_{l}
$$

where $c_{j l}$ is now a matrix of paramters and $\alpha_{l} \in B^{2}\left(P_{0}\right)$ a set of well chosen basis functions defining a space for numerical approximation of the distribution and its dependence on $X$.

The second challenge is that, unlike the Aitchison simplex, the Bayes space does contain functions which are not distributions: without further restrictions, Equation (1) might result in an improper prior. However, there is a simple trick to avoid this. According to van den Boogaart, Egozcue and Pawlowsky-Glahn (2014) the subset of all functions with finite $P_{0}$-density contains only distributions and is dense in $B^{2}\left(P_{0}\right)$. One way to avoid problems with distributions completely would thus be to choose the $\alpha_{l}$ as measures with finite $P_{0}$-density.

### 2.3 Parameter estimation

In this model, the joint distribution of $n$ realisations of the dependent variable is given by an


$$
f(Y \mid X)=\prod_{i=1}^{n} \frac{\exp \left(X_{i}^{t} C^{t} S\left(Y_{i}\right)\right)}{\int \exp \left(X_{i}^{t} C^{t} S(y)\right) d P_{0}(y)}
$$

$$
\begin{aligned}
& =\frac{\exp \left(\sum_{i=1}^{n} \operatorname{tr} S\left(Y_{i}\right) X_{i} C\right)}{\prod_{i=1}^{n} \int \exp \left(X_{i}^{t} C^{t} S(y)\right) d P_{0}(y)} \\
& =\frac{\exp \left(\operatorname{tr}\left(\sum_{i=1}^{n} S\left(Y_{i}\right) X_{i}\right) C\right)}{\prod_{i=1}^{n} \int \exp \left(X_{i}^{t} C^{t} S(y)\right) d P_{0}(y)}
\end{aligned}
$$

where $S(y):=\left(\ln \left(\frac{d \alpha_{j}(y)}{d P_{0}(y)}\right)\right)_{j}$. It is thus an exponential family with sufficient statistic $\sum_{i=1}^{n} S\left(Y_{i}\right) X_{i}$ and the matrix $C=\left(c_{j k}\right)_{j k}$ as natural parameters. As the domain of valid parameters is the whole real space $R^{p \times N}$ the parameter space is open and linked to a regular exponential family (Johansen 1979). Parameters can be estimated by a maximum likelihood estimator, typically efficiently computable through a Fisher scoring algorithm:
$\operatorname{vec}\left(C_{s+1}\right)=\operatorname{vec}\left(C_{s}\right)-\operatorname{var}_{C_{s}}\left(\operatorname{vec}\left(\sum_{i=1}^{n} S\left(Y_{i}\right) X_{i}\right)\right)\left(E_{C_{s}}\left[\operatorname{vec}\left(\sum_{i=1}^{n} S\left(Y_{i}\right) X_{i}\right)\right]-\operatorname{vec}\left(\sum_{i=1}^{n} S\left(Y_{i}\right) X_{i}\right)\right)$
The expected value and the variance can be computed by numerical integration:

$$
\begin{aligned}
c_{i}(C) & :=\int \exp \left(\operatorname{tr} S\left(Y_{i}\right) X_{i} C\right) d P_{0}(y) \\
e_{i}(C) & :=\frac{1}{c_{i}(C)} \int S\left(Y_{i}\right) \exp \left(\operatorname{tr} S\left(Y_{i}\right) X_{i} C\right) d P_{0}(y) \\
E_{C}\left[\operatorname{vec}\left(\sum_{i=1}^{n} S\left(Y_{i}\right) X_{i}\right)\right] & =\sum_{i=1}^{n} e_{i}(C) X_{i} \\
v_{i}(C) & :=\frac{1}{c_{i}(C)} \int S(y)^{t} S(y) \exp \left(\operatorname{tr} S\left(Y_{i}\right) X_{i} C\right) d P_{0}(y)-e_{i}(C)^{t} e_{i}(C) \\
\operatorname{var}_{C}\left(\operatorname{vec}\left(\sum_{i=1}^{n} S\left(Y_{i}\right) X_{i}\right)\right) & =\sum_{i=1}^{n} v_{i}(C) \otimes X_{i}^{t} X_{i}
\end{aligned}
$$

We use a Monte Carlo Method with a fixed sample form $P_{0}$ to do the integrations. Using a fixed sample removes the variablity between the iterations, keeps the numerical error constant between the iterations, and thus allows the deterministic iteration to converge despite the random error of the Monte Carlo integration.

## 3 Numerical example

For a worked example we use R (R Development Core Team 2011). As a reference measure $P_{0}$ we use the uniform distribution on $[0,1]$, which is described by a simulation function:

```
> PO = runif
```

The basis $\alpha_{l}, l=1, \ldots, N$ for the approximation is a user choice. The number of model parameters is proportional to the number of basis elements. The number should thus be chosen dependent on the size of the dataset available. In this first example we use a basis formed by polynomials of degrees 1 to 4 in clr representation (Fig. 1), described by their $P_{0}$-density:

```
> dalpha = function(y) exp(-outer((y-0.5)*2,1:4,"^"))
```

For the simulation of the regression we need independent variables:

```
> n = 1000
> X = cbind(1,rnorm(n),rnorm(n))
```



Figure 1: Basis functions used.
and to simulate a dataset, we require true regression coefficients
> $(C=\operatorname{sign}(\operatorname{matrix}(-\operatorname{rnorm}(n c o l(X) * l e n g t h(d a l p h a(0))), n r o w=n c o l(X))))$

|  | $[, 1]$ | $[, 2]$ | $[, 3]$ | $[, 4]$ |
| :--- | ---: | ---: | ---: | ---: |
| $[1]$, | -1 | 1 | -1 | 1 |
| $[2]$, | 1 | 1 | 1 | -1 |
| $[3]$, | 1 | 1 | -1 | 1 |

To simulate this regression model we have used a simplified rejection sampling algorithm:

```
> sim <- function(fc,n) apply(fc,1,function(fc) {
+ p=exp(log(dalpha(u<-PO(n)))%*%fc);
+ sample(u,1,prob=p/sum(p))
+ })
> Y <- sim(X%*%C,1000)
> par(mfrow=c(1,2))
> plot(Y~X[,2],pch=20);abline(lm(Y~ X[,2]))
> plot(Y~X[,3],pch=20);abline(lm(Y~ X[,3]))
```

The actual estimation routine is quite compact:

```
> ProbRegEst <-function (Y,X,Dalpha=dalpha,PO=PO,CO=matrix(0,p,N),IS=PO(10000),iter=30) {
    T = t(X)%*%log(Dalpha(Y))# Regression Statistics of data
    S = log(Dalpha(IS)) # Statistics of integration sample
    N = ncol(T) # Number of statistics
    p = ncol(X) # Number of parameter vectors
    n = nrow(S) # Integration sample size
    conv=numeric(iter) # Space to store iterations
```



Figure 2: Scatterplots showing the dependency of the random variable $Y$ on the explanatory variables as described by the simulated model. Only a weak correlation is visible.

```
    C = c(C0) # Initial Parameter vector
    for(i in 1:iter) { # loop over iterations
    expCS = exp(X%*%structure(C,dim=dim(CO))%*%t(S)) # scaled conditional densities
    w = expCS/rowSums(expCS) # integration weights
    ES = w%*%S # expected values for each case
    ET = t(X)%*%ES # expected value of overall statistics
    # variance per case:
    VS = (w%*% ((S[,rep (1:N,N)]*S[,rep(1:N,each=N)])))-(ES[,rep(1:N,N)]*ES[,rep(1:N,each=N)])
    # variance of overall Statistics:
    aux = t(X[,rep (1:p,p)]*X[,rep (1:p,each=p)])%*%VS
    VT = structure(aperm(structure(aux, dim=c(p,p,N,N)),c(1,3,2,4)),\operatorname{dim=c}(p*N,p*N))
    conv[i] = sum((T-ET)^2) # convergence criterion
    C = C - solve(VT,c(ET-T)) # Fisher scoring iteration
    }
    structure(C,dim=dim(CO),conv=conv) # Return parameter and convergence
}
```

This command has as parameters: the realisations of the dependent variable $Y$, the design matrix of the Bayes space linear model X containing a constant 1 column and the regressors as columns, the $P_{0}$-density functions of the $\alpha_{l}$ basis elements and a random number generator function P0 for the reference measure $P_{0}$. The following parameters are optional: C0 is the starting parameter value for the Fisher Scoring iteration, initialized to a 0 matrix by default; the fixed intergration sample IS initialized to a Monte Carlo sample by default; and the number of iterations iter.

The actual estimation is done by

```
> (Cest <- ProbRegEst(Y,X,dalpha,PO,iter=10))
```

|  | $[, 1]$ | $[, 2]$ | $[, 3]$ | $[, 4]$ |
| :--- | ---: | ---: | ---: | ---: |
| $[1]$, | -0.8324125 | 0.3455492 | -1.1221907 | 1.5539929 |
| $[2]$, | 0.7547944 | 0.9398062 | 1.1863786 | -1.2216265 |
| $[3]$, | 0.9075868 | 1.9917717 | -0.9443885 | -0.1474773 |



Figure 3: The first seven simulated data points, with their conditional density in the simulation (solid) and the estimated conditional density refitted from the dataset

```
attr(,"conv")
    [1] 2.646810e+05 2.749298e+04 5.069689e+02 2.553945e+01 2.825084e-01
    [6] 3.600114e-05 4.820561e-13 2.249705e-25 2.221432e-26 1.454028e-26
>C
```

|  | $[, 1]$ | $[, 2]$ | $[, 3]$ | $[, 4]$ |
| :--- | ---: | ---: | ---: | ---: |
| $[1]$, | -1 | 1 | -1 | 1 |
| $[2]$, | 1 | 1 | 1 | -1 |
| $[3]$, | 1 | 1 | -1 | 1 |

In this example the algorithm shows a fast convergence in only a few steps, towards a result numerically identical to a maximum likelihood solution. The fitted parameters show a considerable relative error with respect to their true values, which is not a suprise when estimating 12 parameters from only thousand data points. Figure 3 compares true and predicted distributions for the first seven samples.

## 4 Conclusions

Linear models in Bayes spaces provide a natural way of regressing the whole conditional distribution against covariables. The resulting probablity regression can be seen as a generalisation of Multinomial logistic regression. Through the choice of the approximating basis and the models design matrix, the user has considerable influence on the ability of the model to approximate the dependence. The more flexible the models get, the more parameters must be estimated and the more data is needed for a stable estimation. The current estimation approach works efficiently, but not yet very reliably. From time to time the estimator does not converge. This can have different types of reasons. One is that for some data sets of the same model, a maximum likelihood estimator does not exist. The second is that the algorithm might diverge without a step size control.

## References

## References

Lehmann, M (2014). Logistische Regression in Bayes Räumen mit kategoriellen und stetigen Responsevariablen, Master Thesis, TU Bergakademie Freiberg, Institute for Stochastics

McCullagh, P. and Nelder, J. A. ( 1989). Generalized Linear Models, Vol. 37 of Monographs on Statistics and Applied Probability, 2 edn, Chapman and Hall, London.

Johansen, S. (1979). Introduction to the theory of regular exponential families, Institute of Mathematicals Statistics, University of Copenhagen, Lecture Notes 3, Copenhagen.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing, ISBN 3-900051-07-0 http://www.R-project.org/
van den Boogaart, K. G., Egozcue, J. J. and Pawlowsky-Glahn, V. (2014). Bayes Hilbert Spaces, Australian \& New Zealand Journal of Statistics, 56: 171-194. doi: 10.1111/anzs. 12074
van den Boogaart, K. G., Tolosana, R. (2013). Analysing Compositional Data with R, UseR!, Spinger, 257p.
van den Boogaart, K. G., Tolosana-Delgado, R., Lehmann, M., Mueller, U. (2014). On the joint multi point simulation of discrete and continuous geometallurgical parameters Proceedings of SMP 2014, Ore Body modelling and strategic mine planning, 24.-25.11.2014, Perth, Australia

# From univariate background (baseline) values towards the concept of compositional background (baseline) values 

A. Buccianti ${ }^{1}$, B. Nisi $^{2}$, and B. Raco ${ }^{2}$<br>${ }^{1}$ University of Florence (I); antonella.buccianti@unifi.it ${ }^{2}$ IGG-CNR Pisa (I)


#### Abstract

Water geochemistry is often investigated by considering a large number of variables, including major, minor and trace elements. Some of these are usually highly correlated due to coherent geochemical behaviour, but the effect of anthropic factors tends to increase data variability, sometimes obscuring the natural laws governing their relationships. In this framework it may be difficult to identify geochemical features linked to natural phenomena as well as to separate geogenic anomalies from the anthropogenic ones or to define background or baseline concentrations for single chemical elements. This is particularly true at regional level where numerous phenomena may interact and mix together, forming a complex pattern not easy to interpret. The identification of background or baseline values is also complicated due to the compositional nature of chemical variables, so that under the Compositional Data Analysis Theory (CoDA) single background or baseline values lose their meaning. However they are fundamental references for public institutions and government policies. In this contribution a new approach is proposed aimed at investigating the regionalised structure of the geochemical data by considering the joint behaviour of several chemical elements or compositions. The approach is based on the robust CoDA theory so that the proportionality features of abundance data are fully taken into account, enhancing their relative multivariate behaviour, as well as the influence of outliers. An application example is presented for the water compositions sampled in Tuscany Region, a surface of about $23,000 \mathrm{~km}^{2}$, a project sponsored by the regional government.


Key words: water geochemistry; background values; baseline values; compositional data analysis; robust statistics.

## 1 Introduction

Geogenic or 'natural background' substances in the environment are known to occur as concentrations in air, soils and waters that can be hazardous for life forms. As reported in Nordstrom (2015) it may be difficult for water chemistry to refer to 'natural background' as unpolluted or pristine preindustrial conditions. This happens because 1) a widespread global contamination by several trace constituents has occurred, 2) natural variations can be large so that a single analytical result for a given element or compound cannot be useful, 3) the effect of scale and study objectives could have played a very important role in the determination of background (Reimann and Garrett, 2005). One of the most interesting discussions on this subject is reported in Matschullat et al. (2000). These authors recognise that the citation of single values for a geochemical background is neither useful for the identification of the geogenic contribution nor for the determination of an anthropogenic contamination, because single values do not yield information about natural deviation. Another term, geochemical baseline, is often closely associated with geochemical background and they are often used as synonyms (e.g., Nieto et al. 2005 Galuzska, 2007). However, the geochemical baseline has to be considered as the natural background in an area heavy polluted by human activities.

The knowledge of groundwater chemistry and of associated background (baseline) values for elements and compounds is a priority for human health as established by the environmental organisms and institutions of many industrialised and developing countries (World Health Organisation, 1992; UNWater, 2011). Notwithstanding the importance of this item, the definition of the background or baseline content in water is difficult and cannot be based only on a good analytical phase and on a time-limited sampling campaign. In fact, water chemistry can vary in space depending on changing climate conditions as well as on the contribution of anthropogenic pollution. Moreover, in the same place the abundance of chemical species can change in time due to the effect of several environmental and anthropic factors. Consequently it is evident that single concentration values are not able to give representative information about the influence of complex phenomena as occurring in groundwater systems. In this natural reservoir only simultaneous chemical equations are able to describe the different mineral/water equilibriums and the investigation of data variability assumes a fundamental role. For these motifs it is our opinion that only a compositional approach is able to describe the conditions of such a complex phase as the groundwater chemistry (Aitchison, 1986; Buccianti et al., 2014). Moreover, due to the expected presence of anomalous values, robust methods can improve the compositional approach (Rousseeuw, 1984; Maronna and Zamar, 2002; Verboven and Huber, 2005; Daszykowski et al., 2007). Thus the combination of compositional and robust methods applied on water chemistry in anthropic areas permits us to evolve the concept of baseline for single values to that of compositional baseline. The approach can be extended to different scales in groundwater investigation and represents an implementation, moving from the value of single variables to that of compositions when the joint distribution of $D$ variables is considered (Buccianti and Gallo, 2013; De Caritat and Grunsky, 2013).

## 2 Material and methods

### 2.1 Data sources

The geochemical data used in the present study refer to groundwater samples actually stored in the GEOBASI database (Raco et al., 2015) representing the official repository for the geochemical composition of geological media of Tuscany Region (Central Italy). The chemical compositions of 6,808 cases (spring and wells) were considered, 6,435 of which were geo-referenced and checked through the inspection of the original field maps. For those samples where the main composition ( $\mathrm{Ca}, \mathrm{Mg}, \mathrm{Na}, \mathrm{K}$, $\mathrm{HCO}_{3}, \mathrm{SO}_{4}$ and Cl ) was available, the quality of the geochemical data was checked by means of a simple charge balance and only those waters having a percentage deviation $<10$ ( 4,804 samples) were taken into account for further processing. For the sake of clarity it should be noted that all the geochemical data were used, including those repeatedly analysed over time, so that variability could be also affected by seasonality (Fig. 1).

The investigated area corresponds to the Tuscany Regional district and has a surface of about $23,000 \mathrm{~km}^{2}$. Tuscany is bordered to the northern and southern sectors by the Northern Apennine, an orogenic belt formed by the Cretaceous to Miocene compressive phase related to the collision of the European and African plates. The metamorphic Paleozoic basement (e.g. phyllitic to quarzitic and
micaschists rocks to Triassic evaporitic anyhydrites), the Mesozoic and Cenozoic carbonate and evaporitic formations, overlain by flysch series, as well as granite intrusions and volcanic rocks are typical lithologies (Carmignani et al., 2004).


Figure 1: Position of all the water samples for which the chemical composition is stored in the GEOBASI repository. Cases located in red and brown areas correspond to groundwaters located in alluvial and fracture rock aquifers, respectively.

### 2.1 Statistical methodologies

In this contribution our aim is to shift the attention from the investigation of single variables of compositional data set to their joint multivariate behaviour. To achieve this target, the principles of CoDA (Compositional Data Analysis) theory were followed (Aitchison, 1982), implemented by the application of multivariate robust methodologies (Maronna and Zamar, 2002; Daszykowski et al., 2007). Compositional data are vectors of positive values quantitatively describing the contribution of $D$ parts of some whole, which carry only relative information (Aitchison, 1982; Aitchison, 1986). Due to these features, the Euclidean geometrical approach to the statistical analysis of compositions may give misleading results since compositional data pertain to the simplex sample space and not to the real one (Egozcue and Pawlowsky-Glahn, 2006; Buccianti and Magli, 2011; Buccianti, 2013). The simplex sample space is governed by the Aitchison geometry, and has all the properties of a ( $D-1$ ) dimensional Euclidean space (Egozcue and Pawlowsky-Glahn, 2006). To work in these unconstrained conditions, compositions need to be expressed as vectors of values that belong to such a space. In our case the isometric log-ratio (ilr) transformation, proposed by Egozcue et al. (2003) was adopted. Notwithstanding its theoretical advantages and practical properties, its use may be compromised when coordinates have to be interpreted from a geochemical point of view. However if the concept of balance between groups of parts, originated by a sequential binary partition, is considered (Egozcue and Pawlowsky-Glahn, 2005), this path may be highly simplified. In a sequential binary partition in each of the D-1 steps of the procedure the compositional parts are divided into two non-overlapping groups; the resulting $\mathrm{D}-1$ ilr variables represent balances between these groups in $\mathrm{R}^{\mathrm{D}-1}$ :

$$
\begin{equation*}
\operatorname{ilr}_{i}=\sqrt{\frac{r \times s}{r+s}} \log \frac{g\left(c_{+}\right)}{g\left(c_{-}\right)} \tag{1}
\end{equation*}
$$

with $i=1,2, \ldots, D-1$ and where $\mathrm{g}\left(\mathrm{c}_{+}\right)$represents the geometric mean of the $r$ variables of the numerator of the balance, $\mathrm{g}(\mathrm{c}$.) the geometric mean of the $s$ variables of the denominator. The matrix of ilr coordinates was then analysed by robust methods with the aim to identify samples with anomalous behaviour but simultaneously avoiding their effect on the classical estimates. For this purpose, the robust distance of a composition, labelled $R D_{i}$, is used to detect whether it is an outlier composition or not. It is defined as:

$$
\begin{equation*}
R D_{i}=\sqrt{\left(\mathbf{x}_{i}-\widehat{\boldsymbol{\mu}}_{M C D}\right)^{T} \widehat{\boldsymbol{\Sigma}}_{M C D}^{-1}\left(\mathbf{x}_{i}-\widehat{\boldsymbol{\mu}}_{M C D}\right)} \tag{2}
\end{equation*}
$$

with $\hat{\mu}_{M C D}$ and $\hat{\Sigma}_{M C D}$ as the MCD location and scatter estimates. This robust distance is a robustification of the Mahalanobis one where classical mean and empirical covariance matrix are used as estimates of location and scatter. Under the normal assumption, the outlier compositions are those compositions having a robust distance larger than the cut-off value $\sqrt{\chi_{D-1,0.975}^{2}}$. Robust estimates of the compositional centre $\boldsymbol{\mu}$ and scatter matrix $\boldsymbol{\Sigma}$ can be obtained by using the Minimum Covariance Determinant (MCD) estimator (Rousseeuw, 1984; Verboven and Huber, 2005). For each composition the $R D_{i}$ value from the robust centre was determined and the results mapped to highlight the regionalised behaviour of the investigated groundwaters. The mapping of robust Mahalanobis distance by using different methodologies and approaches is able to indicate 1) in which part of the investigated area the pressure toward anomalous behaviours is higher, 2) where the compositions nearest to the barycentre are and 3 ) if spatial continuity is present in limited portions of the territory. Important information about the geochemical processes, their diffusion and influence in the different parts of the investigated areas can be consequently obtained (Filzmoser and Hron, 2008; Filzmoser et al., 2009a, 2009b; Filzmoser et al., 2012).

All the analyses have been performed by using robust routines developed in Matlab and R (Verboven and Hubert, 2005; R Development Core Team, 2015).

## 3 Results and discussion

In the investigation of the main groundwater composition of Tuscany Region the first step was to transform variables by using the isometric log-ratio conversion. To achieve this aim the sequential binary partition proposed by Egozcue and Pawlowsky-Glahn (2005) was adopted where in each of the $D-1$ steps of the procedure the compositional parts were split into two non-overlapping groups; the resulting $D-1$ ilr variables represented balances between these groups in $\mathrm{R}^{\mathrm{D}-1}$. Thus the balances between two groups of parts are orthogonal log-ratio contrasts between geometric means of the selected non-overlapping groups. The sequential binary partition of [Eq. (1)] was chosen so that all the cations were rationed versus (symbol |) all the anions [ $\left.\mathrm{Ca}, \mathrm{Mg}, \mathrm{K}, \mathrm{Na} \mid \mathrm{HCO}_{3}, \mathrm{Cl}, \mathrm{SO}_{4}\right]$, then $[\mathrm{Ca}, \mathrm{Mg} \mid \mathrm{K}, \mathrm{Na}],[\mathrm{Ca} \mid \mathrm{Mg}],[\mathrm{K} \mid \mathrm{Na}]$, $\left[\mathrm{HCO}_{3} \mid \mathrm{Cl}, \mathrm{SO}_{4}\right]$ and finally $\left[\mathrm{Cl} \mid \mathrm{SO}_{4}\right]$. In this conversion the geometric means are central values in each group of parts, their ratio measures the relative weight of each group and the logarithm provides the appropriate scale; the square root coefficient of [Eq. (1)] is a normalising constant which allows the comparison of numerically different balances. A positive balance means that, in (geometric) mean, the group of parts of the numerator has more weight in the composition than the group of the denominator (and conversely for negative balances). The calculus of the robust distance $R D_{i}$ from the barycentre was obtained by using the Minimum Covariance Determinant (MCD) estimator (Verboven and Hubert, 2005). From a statistical point of view it was possible to discriminate 901 anomalous compositions. These under the normal assumption have a robust distance larger than the cut-off value $\sqrt{\chi_{D-1,0.075}^{2}}$ with comparison to the remaining 3543 cases. Their relative position is reported in Figure 2. As we can see, compositions below the statistical threshold (grey points) cover most part of the Region, while composition above this reference point (black points) mainly pertain to the coastal areas (saline intrusion) and to some zones where presence of mineralisation, geothermal activities or pollution due to anthropic contributions can explain the presence of geochemical anomalies. This discrimination is supported by the chemical composition of the sample nearest to the robust barycentre pertaining to a typical $\mathrm{Ca}-\mathrm{HCO}_{3}$ geochemical facies (composition in $\mathrm{mg} / \mathrm{L}$ equal to $\mathrm{HCO}_{3}=581, \mathrm{Ca}=168, \mathrm{Cl}=803, \mathrm{Mg}=32.7, \mathrm{~K}=3.4, \mathrm{Na}=67.9$, $\left.\mathrm{SO}_{4}=97.6\right)$ and by the composition of the samples with the higher distance from the robust barycentre pertaining to a $\mathrm{Ca}-\mathrm{SO}_{4}$ geochemical facies (composition equal to $\mathrm{HCO}_{3}=4.8, \mathrm{Ca}=356, \mathrm{Cl}=270, \mathrm{Mg}=$ $104, \mathrm{~K}=2.4, \mathrm{Na}=250, \mathrm{SO}_{4}=1250$ ).
The histogram of the robust Mahalanobis distance for the data below the statistical threshold ( $n=3543$ ) is reported in Figure 3. Its analysis is important to understand how compositions move from the barycentre. If the robust barycentre is candidate to be a compositional baseline the investigation of variability of compositional changes as monitored by the behaviour of the robust Mahalanobis distance is not a secondary feature.
The $R D_{i}$ data distribution appears to be slightly asymmetrical and apparently classical normal or lognormal models are not able to describe it. A model that leads towards much heavier tails is the power law distribution, often used to describe inhomogeneous and irregular distribution of element concentrations in several geological situations. When a probability density function displays a heavy non-

Gaussian tail, this may be indication of the presence of multifractal processes. This also indicates that the system as a whole is experiencing a non-linear dissipation in the energy interchange among different scales. From this perspective the shape of a probability density function can be a powerful diagnostic tool in environmental problems.
A nonnegative random variable $X$ is said to have a power law distribution if:

$$
\begin{equation*}
\operatorname{Pr}[X \geq x] \cong c x^{\alpha} \tag{3}
\end{equation*}
$$

for constants $\mathrm{c}>0$ and $\alpha>0$, so that asymptotically the tails fall according to the power $\alpha$. For a power law distribution usually $\alpha$ falls in the range $0<\alpha \leq 2$, in which case $X$ has infinite variance. If $\alpha \leq 1$ then $X$ also has infinite mean. The scaling exponent $\alpha$ is usually called the fractal dimension. From a general point of view, the model leads to much heavier tails than other common models explaining the concentration of minor and trace elements in geological materials (Goncalves, 2001; Ma et al. 2014).


Figure 2: Position of the groundwater samples below (grey points) and above (black points) the statistical cut-off value.


Figure 3: Histogram of the Mahalanobis distance of groundwater samples below the cut-off value $\sqrt{\chi_{D-1,0.975}^{2}}$


Figure 4: Cumulative number $N$ of samples whose robust Mahalanobis distance is higher than a given value $R D_{i}$.

An interesting feature of this distribution model is that if $X$ has a power law distribution, then in a log-log plot of $\operatorname{Pr}[X \geq x]$ the pattern of points will be described by a straight line. The graphical analysis is often used to visualize breaks in the data distribution (change in slope and presence of more than one straight line) and to investigate geochemical anomalies versus background (Agterberg, 2014). Results for the studied groundwater data are reported in Figure 4.
As we can see a single straight line is not sufficient to describe all the data. Apparently, compositional changes that move compositions from the robust barycenter are multifractal and the physical-chemical reasons for this behavior could be related to the presence of non-linear interactions between different scales and to the inhomogeneous character of dissipation of the chemical reactions. In general multifractality is a property of a dynamical system in which energy dissipation cannot be neglected. This condition leads to the presence of extended areas (or intervals) of low fluctuations intermittent with small areas of extremely large fluctuations. The spatial distribution of the $R D_{i}$ values is reported in Figure 5.


Figure 5: Spatial distribution of the $R D_{i}$ values. Values increase from yellow $(0.2879)$ to brown $(3.80)($ median value $=2.023)$.

As we can see, compositions similar to that of the robust barycenter (yellow points) are near to or in some cases they overlap compositions very far from this reference (brown points), indicating that the chemistry of our samples can change in a short spatial range.
On the whole, results indicate that the investigated water-rock system developed under conditions far from equilibrium as a progressively self-organizing dissipative structure (Kondepudi and Prigogine, 1998). A "dissipative structure" is a non-equilibrium system, far from an equilibrium state, and should be supported by continuous inputs and outputs of materials in open conditions. Several authors have discussed the features of dissipative structures for groundwater flow systems and human activity appears to be a significant perturbation factor ( Xu and $\mathrm{Du}, 2014$ ). In this respect, the evolution of the investigated groundwater system may be attributable to a dissipative process of macroscopic states being permuted by ones influenced by natural and human factors. Consequently the groundwater system is an open and complicated framework where interactions are governed by non-linear dynamics. In this context the identification of baseline compositions appears to be more representative of single values in the understanding of the processes working in the investigated area.

## 4 Conclusions

The term " geochemical baseline", officially introduced in 1993 in the context of the International Geological Correlation Program (IGCP Project 360), Global Geochemical Baselines , refers to the natural variation in the concentration of an element in the media of the superficial environment. The term can indicate the actual content of an element in the superficial environment matrices at a given point in time. It includes the geogenic natural concentrations (natural background) and the anthropogenic contribution.

Compositional changes from the robust barycenter for groundwater chemistry of Tuscany Region were analyzed by investigating the behavior of the robust Mahalanobis distance from it. The distance was calculated after having transformed original concentrations by using balances (isometric logratio transformation). The aim was to verify if the robust barycenter could be considered a possible compositional baseline. Results indicate that the distance obeys the power law distribution and displays properties of multifractality. In this perspective compositional changes are neither completely deterministic not totally chaotic. Rather they are in an intermediate state, which possesses a property of multifractality in spatial domain and probable intermittency in time. Since fractal structures form spontaneously only in the presence of a complex dissipative structure it is evident that the investigated groundwater system is an open and complicated framework where interactions are governed by non-linear dynamics. Moreover, compositional changes are characterized by self-similarity patterns. This means that the physical-chemical laws that control spatial and temporal variability on one scale also control patterns and spatial variability on other scales, implying scale-independence. All of the previous results, even if preliminary, require caution in the definition of baseline concentrations for single values without the use of a compositional approach and a global conceptual model of the groundwater system for all its interconnected components.

## Acknowledgements

The research was developed thanks to funds of the University of Florence (2014) and of Tuscany Region through the Geobasi project (2009-2014).

## References

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). Journal Royal Statistical Society Ser. B-Stat. Methodol. 44(2), 139-177.
Aitchison, J. (1986). The Statistical Analysis of Compositional Data, Chapman and Hall, London, p. 463.
Buccianti, A. and Magli, R. (2011). Metric concepts and implications in describing compositional changes for world rivers water chemistry. Computer \& Geoscience 37(5), 670-676.
Buccianti, A. (2013). Is compositional data analysis a way to see beyond the illusion? Computers \& geosciences 50, 165-173.

Buccianti, A., Gallo, M. (2013). Weighted principal component analysis for compositional data: application example for the water chemistry of the Arno river (Tuscany, central Italy). Environmetrics 24, 269-277.
Buccianti, A., Egozcue, J.J., Pawlowsky-Glahn, V (2014). Variation diagrams to statistically model the behaviour of geochemical variables: theory and applications. Journal of Hydrology, 519(PA), 988998.

Carmignani L., Conti P., Cornamusini G. \& Meccheri M. (2004): The internal northern Apennines, the northern tyrrhenian sea and the Sardinia-Corsica block, in Geology of Italy. Italian Geological Society Bulletin IGC32 Florence - 2004, 59-77.
Daszykowski M., Kaczmarek, K., Vander Heyden, Y., Walczak, B. (2007). Robust statistic in data analysis. A review. Basic concept. Chemometrics and Intelligent Laboratory System 85, 203-219.
De Caritat, P., and Grunsky, E. (2013). Defining element associations and inferring geological processes from total element concentrations in Australia catchment outlet sediments: Multivariate analysis of continental-scale geochemical data. Applied Geochemistry 33, 104-126.
Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barcelo-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. Mathematical Geology 35(3), 279-300.
Egozcue, J.J. and Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. Mathematical Geology 37(7),795-828.
Egozcue, J.J. and Pawlowsky-Glahn, V. (2006), Simplicial geometry for compositional data, in: Buccianti, A., Mateu-Figueras G. and Pawlowsky-Glahn V. (Eds.) Compositional Data Analysis in the Geosciences: from theory to practice, Special Publication 264, Geological Society, London, 1228.

Filzmoser P., Hron K. (2008). Outlier detection for compositional data using robust methods. Mathematical Geosciences 40(3), 233-248.
Filzmoser, P., Hron, K. and Reimann, C. (2009a), Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Science of the Total Environment 407, 6100-6108.
Filzmoser, P., Hron, K. and Reimann, C. (2009b), Principal component analysis for compositional data with outliers. Environmetrics 20(6), 621-632.P. Filzmoser, K. Hron, and C. Reimann (2012). Interpretation of multivariate outliers for compositional data. Computers \& Geosciences, 39, 77-85.
Galuszka, A. (2007) A review of geochemical background concepts and an example using data from Poland. Environmental Geology 52, 861-870.
Goncalves, M.A. (2001), Characteriszation of geochemical distributions using multifractals models. Mathematical Geosciences 33, 41-61.
Kondepudi, D., Prigogine, I., (1998). Modern Thermodynamics. From heat engines to dissipative structures. John Wiley \& Sons, 486 p.
Ma, T., Li, C. and Lu, Z. (2014), Estimating the average concentration of minor and trace elements in surficial sediments using fractal methods. Journal of Geochemical Exploration 139, 207-216.
Maronna, R. A. and R. H. Zamar (2002). Robust multivariate estimates for highdimensional datasets. Technometrics 44, 307-317.
J. Matschullat, R. Ottenstein, C. Reimann (2000). Geochemical background - can we calculate it? Environmental Earth Sciences, 39(9), 990-1000.
Nieto, P., Custodio, E., Manzano, M. (2005) Baseline groundwater quality: a European approach. Environmental Science Policy 8, 399-409
Nordstrom, D.K. (2015). Baseline and premining geochemical characterization of mined sites. Applied Geochemistry doi:10.1016/j.apgeochem.2014.12.010.
R Development Core Team (2015). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0. URL http://www.Rproject.org/.
Raco, B., Buccianti, A., Corongiu, M., Lavorini, G., Macera, P., Manetti, F., Mari, R., Masetti, G., Menichetti, S., Nisi, B., Protano, G., Romanelli, S. (2015). The geochemical database of Tuscany Region (Italy). Italian Journal of Groundwater AS12055: 007 - 018, DOI 10.7343/AS-100-15-0127.
Reimann, C., Garrett, R.G. (2005). Geochemical background - concept and reality. The Science of the Total Environment 350, 12-27.
Rousseeuw, P.J. (1984). Least median of squares regression. Journal American Statistical Association 79, 871-880.
UN-Water (2011). Annual report, 32 p., http://www.unwater.org/.
Verboven, S., and Hubert, M. (2005). LIBRA: a Matlab libray for robust analysis. Chemometrics and Intelligent Laboratory Systems 75, 127-136.

World Health Organisation (WHO) (1992). Our plane, our health. Report of WHO Commission on Health and Environment, Geneva, World Health Organisaton.
Xu, W., Du S. (2014) Information entropy evolution for groundwater flow system: a case study of artificial recharge in Shijiazhuang city, China. Entropy 16, 4408-4419.

# MANOVA of Compositional Data with a Total 

G. Coenders ${ }^{1}$, B. Ferrer-Rosell ${ }^{2}$, G. Mateu-Figueras ${ }^{2}$, and V. Pawlowsky-Glahn ${ }^{2}$<br>${ }^{1}$ University of Girona, Spain; germa.coenders@udg.edu<br>${ }^{2}$ University of Girona, Spain


#### Abstract

Microeconometric analysis of tourist expenditure as a function of traveller characteristics has focused on total absolute expenditure, on absolute expenditures per trip budget parts, and in relative terms per trip budget parts. The analysis of absolute trip expenditure by budget parts confounds how much (budget volume) and how (budget allocation) tourists spend, while that of trip budget share ignores budget volume. In this paper, using tools of compositional data analysis with a total and MANOVA, we bring trip budget share and trip total expenditure together. Multivariate MANOVA tests are invariant to the choice of total, which may be e.g. the log geometric mean of any number of absolute expenditures of components. We show how trip budget components can be combined with a choice of total expenditure alternatives, according to the researchers' questions. We provide an illustration with low cost airline users travelling to Spain. Results show that some variables affect only the expenditure allocation, some others affect only the total, and some affect both: how and how much tourists spend.


Kew words: CODA with a total; tourist expenditure; trip budget share; isometric logratio transformation; MANOVA

## 1 Introduction

Microeconometric analysis of expenditure made by individual tourists as a function of traveller characteristics has focused on a single expenditure variable, on absolute expenditures per trip budget parts, and on expenditure in relative terms (share) per trip budget parts. The vast majority of microeconomic tourism demand studies ( 24 out of 27 in the review of Wang and Davidson, 2010) concern the prediction of one single aggregated expenditure variable.

Another stream of research is that which analyses tourist expenditure per trip budget part (e.g. lodging, food, transportation and sightseeing/entertainment). A common argument for studying tourist expenditure per budget part is that it provides vital information to travel organizers and destination marketers. Not all budget parts have the same economic impact on each destination or on each tourism industry. Besides, knowing expenditure by tourist products is useful when designing the appropriate marketing strategies. For this purpose, researchers have used several methods, such as MANOVA (multivariate analysis of variance) for qualitative predictors, and MANCOVA (multivariate analysis of covariance) or SUR (seemingly unrelated linear regressions) for quantitative predictors (Lehto and others, 2001; Marcussen, 2011; Wang and others, 2006). Since part expenditure in absolute terms is related to total expenditure, a common finding in these studies is that some of the explanatory variables affect all budget elements about equally. For example, Marcussen (2011) found that hotel accommodation, eating in restaurants and staying at the capital city led to higher expenditure in all budget parts; and Lehto and others (2001) found that certain countries of origin had higher expenditures in all budget parts.

Such results with explanatory variables affecting all budget parts about equally will never be obtained when analyzing expenditure allocation in relative terms, in other words, as budget share (proportion of the trip budget devoted to each budget part). By definition, budget share devoted to one budget part can only increase if share devoted to at least one other part decreases. A further argument for studying share is that a large absolute expenditure on a given trip budget part may either correspond to a tourist with a large overall budget or to a tourist who devotes a large share of the trip budget to that part. The study of trip budget share isolates the effects on how a given trip budget is distributed and results are more clearly interpretable in terms of budget allocation. On the negative side, information about total expenditure (how much they spend) is lost.

The empirical analysis of budget share commonly implies estimating an almost ideal demand system of equations (Deaton and Muellbauer, 1980). This system directly fits budget share as dependent variables in a set of simultaneous regressions (Coenen and van Eekeren, 2003; Fleischer and others, 2011).

The almost ideal demand system has a serious methodological drawback. Let $x_{i l}, x_{i 2}, \ldots, x_{i D}$, be the absolute expenditures of individual $i$ in each of the $D$ budget parts. Compared to absolute data, budget share lies in a constrained $D-1$ dimensional space. A $D$-term budget share measured on individual $i, z_{i l}, z_{i 2}, \ldots, z_{i D}$ has the following constraints:

$$
\begin{equation*}
0 \leq z_{i d} \leq 1 \text { and } \sum_{d=1}^{D} z_{i d}=1, \tag{1}
\end{equation*}
$$

and conforms to what the literature refers to as compositional data. Aitchison (1986) and PawlowskyGlahn and Buccianti (2011) warn against the serious problems that arise when using standard statistical analysis tools on compositional data. Compositional data are non-normal and heteroskedastic. The presence of an error term with an unbounded distribution (usually normal), results in a non-zero probability that actual share lies outside the [0,1] interval (Fry, 2011). In other words, the in fact bounded distribution of budget share results in a misspecification of the almost ideal demand system and of any model fitting percentage share with an unbounded error distribution. Aitchison (1986) started the compositional data analysis (CODA) log-ratio approach to tailor statistical analysis to this constrained space and overcome this limitation. The CODA methodology has since been applied to family budgets (Fry and others, 1996; McLaren and others, 1995) and to trip budgets (Ferrer-Rosell and others, 2015; in press).

Even if specialized CODA techniques are available (e.g. Ronning, 1992; Thió-Henestrosa and MartínFernández, 2005), the easy way involves transforming compositional data so that they can be subject to standard and well-understood statistical techniques (Aitchison, 1986; Ferrer-Rosell and others, 2015; in
press; Fry and others, 1996; McLaren and others, 1995). In short, this implies using the transformed share by means of logarithms of ratios, instead of the raw share. Log-ratios recover the full unconstrained $-\infty$ to $\infty$ range ${ }^{1}$.

The analyses of absolute trip expenditure by budget parts and of trip budget share both have limitations. The first confounds budget volume (how much) and budget allocation (how) while the second ignores budget volume altogether. The analyses of trip budget share and of overall expenditure volume provide answers to different though complementary research questions and managerial purposes. In this paper we propose to predict budget share and total budget volume within the same MANOVA or MANCOVA model. The approach we follow is an extension of that of Pawlowsky-Glahn and others (in press) of adding a total to the log-ratios. Once share and total volume have been transformed, the analysis is no more complicated than standard MANOVA or MANCOVA, and any standard software can be used.

The paper is structured as follows. We first review some of the most common log-ratio transformations. We then present the way how a total is combined with the log-ratios in Pawlowsky-Glahn and others (in press), we suggest new possibilities of such combination, and derive the properties of a MANOVA or MANCOVA model with such combinations. We finally present a MANOVA illustration of trip budgets of low cost airline users.

## 2 Log-ratio transformations

Several log-ratio transformations have been suggested in the early CODA literature (Aitchison, 1982; Egozcue and others, 2003). Any log-ratio may be computed either from absolute expenditure or from share.

The centred log-ratio transformation (clr) computes the log-ratios of each component over the geometric mean of all the components, including itself.

$$
\begin{equation*}
y_{i d}=\ln \left(\frac{z_{i d}}{\sqrt[D]{z_{i 1} z_{i 2} z_{i 3} \cdots z_{i D}}}\right)=\ln \left(\frac{x_{i d}}{\sqrt[D]{x_{i 1} x_{i 2} x_{i 3} \cdots x_{i D}}}\right) \text { with } d=1,2,3, \ldots, D . \tag{2}
\end{equation*}
$$

The clr transformation makes Euclidean distances in the transformed space meaningful. Thus, the clr transformation is commonly used for statistical techniques which are based on a metric, such a cluster analysis. On the negative side, it must be noted that in the clr no dimension is lost: one log ratio is a linear combination of the remaining. Thus, the clr is not appropriate for modelling because it leads to a singular covariance matrix.

Egozcue and others (2003) and Egozcue and Pawlowsky-Glahn (2005) introduced the isometric log-ratio transformation (ilr) which preserves distances in the transformed space while using only $D-1$ log-ratios, so that covariance matrices can be inverted. It can thus be used in virtually all standard statistical analyses based on the assumption that data are realisations of real random vectors, which support the real space endowed with the usual Euclidean geometry (Mateu-Figueras and others, 2011). Contrary to the clr, this alternative transformation is more flexible in that the denominator is not the same in all log-ratios. For this reason, ilr log-ratios can easily be tailored to economically meaningful research questions and hypotheses. Unlike the clr, the ilr does not have a unique algebraic expression but its computation will depend on each case. Certainly, the ilr can be viewed as the coordinates with respect to an orthonormal basis in our constrained space. Thus each basis yields different ilr coordinates (Egozcue and others, 2003). The next section gives the requirements an ilr transformation has to fulfil, and the illustration provides guidelines to compute it.

## 3 MANOVA and MANCOVA of log-ratios with a total

Pawlowsky-Glahn and others (in press) study the properties of the space defined by a composition and a total and conclude that the clr or the ilr transformation together with a total computed as $\sqrt{ } D$ times the

[^0]logarithm of the geometric mean of all absolute values makes distances in the combined space meaningful.

The use of logarithms is attractive for economic analysis of expenditure for a number of reasons (Thrane, 2014) and becomes a must when it is to be combined with share. Pawlowsky-Glahn and others (in press) show that the space of logarithms of absolute values by parts, which is commonly used in economic analysis of expenditure by budget parts, [see Eq. (3)] leads to the same distances as the space generated by Equation (4):

$$
\begin{gather*}
\ln \left(x_{i l}\right), \ln \left(x_{i 2}\right), \ldots, \ln \left(x_{i D}\right) ; \\
\ln \left(\frac{x_{i 1}}{\sqrt[D]{x_{i 1} x_{i 2} \cdots x_{i D}}}\right), \ln \left(\frac{x_{i 2}}{\sqrt[D]{x_{i 1} x_{i 2} \cdots x_{i D}}}\right), \cdots, \ln \left(\frac{x_{i D}}{\sqrt[D]{x_{i 1} x_{i 2} \cdots x_{i D}}}\right), \sqrt{D} \ln \left(\sqrt[D]{x_{i 1} x_{i 2} \cdots x_{i D}}\right) . \tag{3}
\end{gather*}
$$

Pawlowsky-Glahn and others (in press) already pointed at the fact that alternative totals may make sense for particular purposes, once their statistical properties are derived. This paper shows that for MANOVA and MANCOVA, modelling ilr log-ratios and the logarithm of the raw geometric mean (without multiplying by the rather hard-to-interpret constant $\sqrt{ } D$ ) of the absolute values of any number of components is equivalent to modelling the logarithms in Equation (3). This is so because MANOVA and MANCOVA results are preserved under change of basis (Mateu-Figueras and others, 2013).

This makes it possible to build tailor made log-ratios which are interpretable with respect to particular research objectives or questions (e.g. a log ratio of the geometric mean of at-destination expenditures over transportation expenditure to find the drivers of the allocation of the budget parts which are spent at destination) and to focus on tailor-made total absolute expenditure(s) which are more relevant to the problem at hand (e.g. absolute expenditure in activities within a research on active tourism).

If log-ratios and total are chosen following the guidelines above, the transformed data ( $y_{i 1}$ to $y_{i D-1}$ logratios and $t_{i}$ total) and the logarithms of absolute expenditure [Eq. (3)]. are related through the $\Psi$ matrix:

$$
\left(\begin{array}{c}
y_{i 1}  \tag{5}\\
y_{i 2} \\
\vdots \\
y_{i D-1} \\
t_{i}
\end{array}\right)=\boldsymbol{\Psi}\left(\begin{array}{c}
\ln \left(x_{i 1}\right) \\
\ln \left(x_{i 2}\right) \\
\vdots \\
\ln \left(x_{i D-1}\right) \\
\ln \left(x_{i D}\right)
\end{array}\right)
$$

As an example we use the same transformation we will use in the illustration section. We have three expenditure components and we compute the total as the logarithm of the geometric mean of the last two. The transformed data are the following function of the logarithms of absolute expenditure:

$$
\left(\begin{array}{c}
y_{i 1}  \tag{6}\\
y_{i 2} \\
t_{i}
\end{array}\right)=\left(\begin{array}{ccc}
\sqrt{\frac{2}{3}} & -\sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{6}} \\
0 & \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \\
0 & \frac{1}{2} & \frac{1}{2}
\end{array}\right)\left(\begin{array}{l}
\ln \left(x_{i 1}\right) \\
\ln \left(x_{i 2}\right) \\
\ln \left(x_{i 3}\right)
\end{array}\right)
$$

The first $D-1$ rows of $\Psi$ fulfil the requirements of an ilr transformation: they have zero sum, they have unit sums of squares, and each pair of rows among the first $D-1$ have zero scalar products. The last row only contains zero values and one or more equal positive values adding up to 1 . These are sufficient conditions for $\Psi$ being a matrix associated with the isomorphism between the space in Equation (3) and the space generated by Equation (4).

MANOVA and MANCOVA multivariate tests and statistics (e.g., Pillai's trace, Hotelling's trace or Wilk's Lambda) are invariant to how ilr log-ratios are computed and to the choice of which components
are in the geometric mean which is used as total. These tests answer the question whether an explanatory variable is relevant for tourist expenditure as a whole.

Univariate tests referring to the total and to each particular log-ratio are not invariant to the choice. If the log-ratios and total have been selected with respect to particular research questions, these tests can be directly related to them. Hence the importance of the interpretability of each log-ratio and of the selected total.

If interpretable log-ratios and totals are built, separate ANOVA and univariate regression models for each log-ratio and for the total will yield the same results as regards the univariate tests (Thrane, in press). The main strength of the multivariate approach is the multivariate tests. When the number of variables is large, which is often the case in tourist expenditure research (Thrane, 2014), the number of univariate tests grows large and leads to type-I risk inflation and to capitalization on chance. A small number of multivariate tests reduce error inflation: univariate tests are only interpreted for variables which are statistically significant according to the multivariate tests.

## 4 Zero expenditure

If the $x_{i d}$ variables contain zeros, neither logarithms nor log-ratios can be computed. An obvious initial procedure to reduce zeros is to amalgamate small and conceptually similar components with many zeros into larger ones. In tourism budget research it can be useful to group together all expenditure on activities or all expenditure on food, for instance. Nevertheless, one must be aware that amalgamation is a decision which prevents the separate analysis of the amalgamated components, as the usual addition is a non-linear operation in the geometry of the simplex.

In certain instances, some zero components result from individual characteristics, which are called essential zeros in the CODA literature (Aitchison, 1986). Another typology of zeros encountered in the CODA literature is the rounding zero, that is, a component which is present but is too small to be detected by the measurement instrument. This is typical in chemical, biological and geological compositions, and the CODA literature offers ample instruments to deal with rounding zeros.

A classic essential zero example in economics is in household budget research when measuring expenditure on tobacco; it will essentially be zero if all members are non-smokers. In many instances it is not clear whether zero expenditures come closer to being essential or rounding zeros. Tourists may spend a certain amount on activities and shopping on certain trips, but not on others and so surveys of only one trip will unavoidably contain some zeros of this type. Tourists may also forget or fail to report trivial expenses, like post-card shopping, local bus tickets, going to a museum, and the like (Legohérel, 1998). Fry and others (2000) claim that in both situations zeros can be proxied by a value below the smallest non-zero expenditure, and thus be treated as rounding zeros. Simulations show this method performs well if the proportion of zeros is below $10 \%$ (Martín-Fernández and others, 2011).

In the illustration, zeros were present only in one budget category (activities and shopping, $4.7 \%$ zeros). The minimum amount spent by the non-zero group was 2 euro. Zeros were replaced with 1.30 euro, which roughly corresponds to the price of a city bus ticket, the entrance to a subsidized local museum, or a cheap souvenir. Given the small proportion of zeros in this category, no distortion is expected due to the substitution.

## 5 Illustration

### 5.1 Chosen log-ratios and total

In this illustration we consider the same ilr transformation as Ferrer-Rosell and others (in press), with emphasis on interpretability with respect to the researchers' questions or hypotheses.

In general, an interpretable log-ratio transformation is easy to compute whenever there is an interpretable sequential binary partition of components into pairs of groups of components, according to the researchers' objectives (Pawlowsky-Glahn and Egozcue, 2011). These partitions start by dividing components into two clusters and then continue by subdividing one of the clusters into two until each
component constitutes its own cluster. $D$ components always involve $D-1$ partitions. These partitions are best understood with a tree diagram.

A meaningful log-ratio transformation starts by taking ratios of the geometric means of the two component clusters at each partition. Numerators and denominators are interchangeable. In our illustration we consider $x_{i 1}=$ transportation expenditure, $x_{i 2}=$ accommodation and food, $x_{i 3}=$ activities and shopping. An interpretable sequential partition is shown in Figure 1.


Figure 1: Sequential partition of budget parts in our illustration.
The sequential partition in Figure 1 implies that the research questions involve the allocation of total expenditure between transportation and at-destination expenditure and the allocation of at-destination expenditure.

The first log-ratio compares transportation expenditure with the geometric mean of accommodation and food on the one hand and activities and shopping on the other. This ratio is used to observe the share of transportation compared to at-destination expenses. Larger values show a higher relative importance of transportation expenses.

$$
\begin{equation*}
y_{i 1}=\ln \left(\frac{x_{i 1}}{\sqrt{x_{i 2} x_{i 3}}}\right)=\ln \left(x_{i 1}\right)-\frac{1}{2} \ln \left(x_{i 2}\right)-\frac{1}{2} \ln \left(x_{i 3}\right) . \tag{7}
\end{equation*}
$$

The second log-ratio is a ratio of accommodation and food over activities and shopping. This ratio indicates how tourists allocate at-destination expenses.

$$
\begin{equation*}
y_{i 2}=\ln \left(\frac{x_{i 2}}{x_{i 3}}\right)=\ln \left(x_{i 2}\right)-\ln \left(x_{i 3}\right) . \tag{8}
\end{equation*}
$$

As regards the total, we focus on studying the economic impact at destination and we choose the logarithm of the geometric average of the two absolute at-destination expenditures:

$$
\begin{equation*}
t_{i}=\ln \left(\sqrt{x_{i 2} x_{i 3}}\right)=\frac{1}{2} \ln \left(x_{i 2}\right)+\frac{1}{2} \ln \left(x_{i 3}\right) . \tag{9}
\end{equation*}
$$

The choice of the relevant total is as dependent on objectives and research questions as the choice of logratios. For instance, if we would have focused on the impact on the hospitality industry we could just have chosen $\ln \left(x_{i 2}\right)$.

These log-ratios and total imply the following $\Psi$ matrix in Equation (5), which fulfils all sufficient conditions but the unit sum of squares in the first $D-1$ rows:

$$
\Psi=\left(\begin{array}{ccc}
1 & -1 / 2 & -1 / 2  \tag{10}\\
0 & 1 & -1 \\
0 & 1 / 2 & 1 / 2
\end{array}\right)
$$

Hence, each of the $D-1$ first rows is multiplied by the appropriate constants, from which the matrix expression in Equation (6) is obtained. The final log-ratios are:

$$
\begin{gather*}
y_{i 1}=\sqrt{\frac{2}{3}}\left(\ln \left(\frac{x_{i 1}}{\sqrt{x_{i 2} x_{i 3}}}\right)\right)=\sqrt{\frac{2}{3}} \ln \left(x_{i 1}\right)-\sqrt{\frac{1}{6}} \ln \left(x_{i 2}\right)-\sqrt{\frac{1}{6}} \ln \left(x_{i 3}\right) ;  \tag{11}\\
y_{i 2}=\sqrt{\frac{1}{2}}\left(\ln \left(\frac{x_{i 2}}{x_{i 3}}\right)\right)=\sqrt{\frac{1}{2}} \ln \left(x_{i 2}\right)-\sqrt{\frac{1}{2}} \ln \left(x_{i 3}\right) . \tag{12}
\end{gather*}
$$

### 5.2 Sample and variables

In this illustration we use secondary official statistics data from the Encuesta de Gasto Turístico (EGATUR) conducted by the Instituto de Estudios Turísticos (IET), an official agency of the Spanish Ministry of Industry, Energy and Tourism. The EGATUR survey asks about tourist expenditure, and trip and traveller characteristics. It was conducted in 27 major Spanish airports in 2012, by computer assisted personal interview when tourists were about to leave the country. The sample is non-proportionally stratified by country of residence, airport and month. All expenditure variables are provided per person (not per travel group).

The universe in this paper is a subset of the EGATUR universe which consists of European leisure visitors arriving by low cost airlines and spending between one and 120 nights in Spain. Flights from outside Europe are excluded because these airlines mostly operate short-haul flights. The study is centred on only those trips with one single destination, thus excluding multi-stage trips. Tourists who have essential zeros in accommodation (tourists who own a house at the destination, or tourists who stay with friends or relatives), and tourists who do not decide how much they spend on certain components (business and study trips, package tourists, trips paid for by the company, gifts or contests) are also excluded. The final sample size is $n=14,446$.

The following traveller characteristics are used as predictors: travel group, country of residence, education, self-reported income category, being repeat or first-time visitor, gender, age, and professional status. Since all variables are qualitative, we use MANOVA.

### 5.3 Results

Answering the question whether explanatory variables are relevant for tourist expenditure as a whole, gender is the only variable which is not globally significant ( p -value $>0.01$ in the first column of Table 1). For this variable no further tests are interpreted.

On the other hand, regarding univariate tests referring to the chosen total (how much) and to each particular log-ratio (how), we found that education is not significantly related to the total expenditure made at destination; that income category is not significantly related to $y_{2}$ (i.e., to the allocation of atdestination budget between accommodation-food and activities-shopping); and that having or having not been to Spain before does not significantly predict either $y_{1}$ or $y_{2}$ ( p -value $>0.01$ in second to fourth columns in Table 1).

Table 1. Hypothesis test p-values

| Variable | Multivariate <br> Pillai's trace test |  | Univariate F tests |  |
| :--- | :---: | :---: | :---: | :---: |
|  |  | $y_{1}:$ log-ratio <br> transportation <br> /at-destination | $y_{2}:$ log-ratio <br> accommodation <br> /activities | $t: \log$ geometric mean <br> accommodation <br> activities |
| Travel group | 0.000 | 0.000 | 0.000 | 0.000 |
| Country of residence | 0.000 | 0.000 | 0.000 | 0.000 |
| Education | 0.000 | 0.000 | 0.000 | 0.028 |
| Reported income category | 0.000 | 0.000 | 0.192 | 0.000 |
| Repeat visitor | 0.000 | 0.825 | 0.051 | 0.000 |
| Gender | 0.190 | 0.030 | 0.169 | 0.114 |
| Age | 0.000 | 0.000 | 0.000 | 0.000 |
| Professional status | 0.000 | 0.000 | 0.000 | 0.000 |

Table 2. Category frequencies and parameter estimates

| Variable | Parameter | Category Freq. | $\begin{gathered} y_{1}: \text { log-ratio } \\ \text { transportation } \\ \text { /at-destination } \end{gathered}$ | $y_{2}$ : log-ratio accommodation /activities | $t: \log$ geometric mean accommodation activities |
| :---: | :---: | :---: | :---: | :---: | :---: |
|  | Intercept |  | 0.057 | 1.690** | 4.975** |
| Travel group | Alone | 15.7\% | -0.077** | 0.026 | -0.077** |
|  | With family | 17.0\% | 0.103** | -0.111** | 0.103 |
|  | With friends | 19.8\% | -0.083** | -0.238** | -0.083** |
|  | With partner | 47.5\% | 0 | 0 | 0 |
| $\begin{aligned} & \text { Country } \\ & \text { of } \\ & \text { residence } \end{aligned}$ | Germany | 10.3\% | -0.127** | -0.117** | 0.234** |
|  | Italy | 17.4\% | -0.265** | -0.200** | -0.044 |
|  | France | 8.7\% | -0.157** | -0.229** | 0.027 |
|  | Benelux | 11.4\% | -0.053* | -0.133** | 0.043 |
|  | Scandinavia | 11.0\% | -0.004 | -0.221** | 0.363** |
|  | Austria, Swit., Liech. | 3.3\% | -0.094* | -0.287** | 0.284** |
|  | Other | 5.7\% | -0.203** | -0.259** | 0.025 |
|  | UK \& Ireland | 32.2\% | 0 | 0 | 0 |
| Education | Up to high school | 28.2\% | 0.068** | 0.106** | 0.034 |
|  | University degree | 71.8\% | 0 | 0 | 0 |
| Reported income category | High income | 23.7\% | -0.098** | 0.016 | 0.259** |
|  | Low income | 3.0\% | 0.001 | -0.072 | 0.118* |
|  | Medium income | 73.3\% | 0 | 0 | 0 |
| Repeat visitor | Yes | 23.0\% | -0.003 | -0.036 | -0.077** |
|  | No | 77.0\% | 0 | 0 | 0 |
| Gender | Female | 46.7\% | 0.025 | 0.021 | -0.021 |
|  | Male | 53.3\% | 0 | 0 | 0 |
| Age | +45 retired | 6.1\% | -0.070** | 0.192** | 0.193** |
|  | $15-24$ | $11.7 \%$ | -0.109** | -0.171** | 0.148** |
|  | +45 not retired | $28.5 \%$ | -0.008 | 0.057* | 0.094** |
|  | $25-44$ | 53.6\% | 0 | 0 | 0 |
| Prof. status | Homemaker | 2.1\% | 0.186** | 0.295** | -0.101 |
|  | Unemployed | 2.8\% | 0.041 | 0.165** | -0.218** |
|  | Low level employee | 4.9\% | 0.219** | 0.236** | -0.352** |
|  | Students | 6.6\% | 0.161** | 0.139* | -0.355** |
|  | High level employee | 9.7\% | 0.050 | 0.091* | 0.054 |
|  | Self employed | 14.2\% | 0.024 | 0.019 | -0.002 |
|  | Mid level employee | 59.7\% | 0 | 0 | 0 |

[^1]As examples of how the results of this type of analysis can be interpreted, we provide one example affecting expenditure allocation (how), one affecting expenditure volume (how much) and one affecting both (Table 2).

Education does not affect how much tourists spend at destination. However, this variable does affect how tourists distribute their trip budget. Tourists with up to secondary education spend comparatively less at destination, and within the expenditures done at destination, they spend comparatively less on activities and shopping, and more on accommodation and food.

Tourists repeating visit to Spain spend absolutely less at destination than those who are for first time in Spain. No differences are found between first-timers and repeat visitors in how they allocate their trip budget.

Age affects both how and how much tourists spend at destination. Regarding trip budget allocation, retired tourists spend comparatively more at destination and less on transportation, and within atdestination expenses, they spend relatively more on accommodation and food. While young (15-24 years’ old) travellers also spend comparatively less on transportation, within at-destination expenses they spend comparatively more on activities and shopping. They are the age group who spends comparatively the most on activities. As far as the at-destination expenditure volume is concerned, old and young tourists spend more than medium-age travellers.

## 6 Conclusions and discussion

Expenditure allocation and expenditure volume are both interesting to tourism scholars and managers alike. In this paper we have combined compositional analysis of trip budget share with the analysis of a trip budget total and have shown how log-ratios and totals can be tailored to the research questions.

The approach in this paper avoids confounding how tourists distribute their expenditure and how much they spend. Table 1 shows which variables affect expenditure allocation and which affect total expenditure. On the contrary, the common modelling of the log absolute expenditures in each budget part, usually leads to concluding that certain tourist characteristics affect all budget parts about equally. When running the latter analysis with the same data, results show that high income earners, tourists travelling alone, retired tourists, and tourists coming from Germany, Scandinavia, Austria, Switzerland and Liechtenstein spend more in all budget parts, which hides the effects on budget allocation. The suggested approach is basically a rotation of the common approach by means of a change of basis. As such it yields identical multivariate tests, while leading to more interpretable univariate tests and estimates. It also shares its statistical assumptions, which have to be checked for the particular data set at hand by means of the usual diagnostics in MANOVA and MANCOVA models.

The appeal of the CODA methodology for studying tourism budgets lies in the fact that, once the variables have been transformed, the researcher can use standard and well understood statistical models. The CODA methodology offers the potential to construct tailor-made log-ratios which are intuitive to interpret and suit the research questions at hand. A partition tree of components is a clear and useful tool in this respect. The CODA methodology has been rightly criticised for ignoring the total even when it is available. Drawing from Pawlowsky-Glahn and others (in press) this paper shows alternative ways to include the total in the analysis, which can also be tailored to the research questions.

Zero expenditure has deserved wide attention both in the econometric literature and in the standard CODA literature. Further research is needed as to zero replacement or analysis when compositions are combined with a total which is known and of interest. The illustration contained in this paper is of course limited by the available variables in the EGATUR survey. Further research is needed with a comprehensive list of predictors (Thrane, 2014).

## Acknowledgements

The authors are pleased to acknowledge the support of the Spanish Institute of Tourism Studies (Instituto de Estudios Turísticos) in providing them with the raw data. The authors have all responsibility for any lack of accuracy or reliability of their data analysis. All authors want to thank Dr. Neelu Seetaram for her
comments to previous versions of the paper and to acknowledge the support of the Catalan Autonomous Government Consolidated Research Group Grant 2014SGR551 funding their research group "Compositional and Spatial Data Analysis (COSDA)". The first two authors want to acknowledge the support of the Spanish Health Ministry grant CB06/02/1002 funding their research group "CIBER of Epidemiology and Public Health (CIBERESP)". The last two authors want to acknowledge the support of the Spanish Economy and Competitiveness Ministry grant MTM2012-33236 funding their project "METRICS".

## References

Aitchison, J. (1982). The statistical analysis of compositional data. Journal of the Royal Statistical Society. Series B (Methodological) 44(2), 139-177.

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. London: Chapman and Hall.

Bewley, R. (1982). The generalised addilog demand system applied to Australian time series and cross section data. Australian Economic Papers 21(38), 177-192.

Coenen, M., and L. van Eekeren (2003). A study of the demand for domestic tourism by Swedish households using a two-staged budgeting model. Scandinavian Journal of Hospitality and Tourism 3(2), 114-133.

Deaton, A., and J. Muellbauer (1980). An almost ideal demand system. American Economic Review 70(3), 312-326.

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

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

Ferrer-Rosell, B., G. Coenders, and E. Martínez-Garcia (2015). Determinants in tourist expenditure composition: The role of airline types. Tourism Economics 21(1), 9-32.

Ferrer-Rosell, B., G. Coenders, and E. Martínez-Garcia (in press). Segmentation by tourist expenditure composition. An approach with compositional data analysis and latent classes. Tourism Analysis.

Fleischer, A., G. Peleg, and J. Rivlin, (2011). The impact of changes in household vacation expenditures on the travel and hospitality industries. Tourism Management 32(4), 815-821.

Fry, T. (2011). Applications in economics. In V. Pawlowsky-Glahn and M. Buccianti (Eds.) Compositional Data Analysis. Theory and Applications, pp. 318-326. New York: Wiley.

Fry, J.M., T.R.L. Fry, and K.R. McLaren (1996). The stochastic specification of demand share equations: restricting budget share to the unit simplex. Journal of Econometrics 73(2), 377-385.

Fry, J.M., T.R.L. Fry, and K.R. McLaren (2000). Compositional data analysis and zeros in micro-data. Applied Economics 32(8), 953-959.

Houthakker, H.S. (1960). Additive preferences. Econometrica 28(2), 244-256.

Legohérel, P. (1998). Toward a market segmentation of the tourism trade: expenditure levels and consumer behaviour instability. Journal of Travel and Tourism Marketing 7(3), 19-39.

Lehto, X.Y., A.M. Morrison, and J.T. O'Leary (2001). Does the visiting friends and relatives' typology make a difference? A study of the international VFR market to the United States. Journal of Travel Research 40(2), 201-212.

Marcussen C.H. (2011). Determinants of tourist spending in cross-sectional studies and at Danish destinations. Tourism Economics 17(4), 833-855.

Martín-Fernández, J.A., J. Palarea-Albaladejo, and R.A. Olea (2011). Dealing with zeros. In V. Pawlowsky-Glahn and M. Buccianti (Eds.) Compositional Data Analysis. Theory and Applications, pp. 47-62. New York: Wiley.

Mateu-Figueras, G., J. Daunis-i-Estadella, and J.A. Martín-Fernández (2013). Analysing grouped CoDa: Manova and Post-hoc contrasts. Paper presented at the Workshop on Compositional Data Analysis (CoDaWork 2013), Vorau.

Mateu-Figueras, G., V. Pawlowsky-Glahn, and J.J. Egozcue, (2011). The principle of working on coordinates. In V. Pawlowsky-Glahn and M. Buccianti (Eds.) Compositional Data Analysis. Theory and Applications, pp. 31-42. New York: Wiley.

McLaren K., J. Fry, and T. Fry (1995). A simple nested test of the almost ideal demand system. Empirical economics 20(1), 149-161.

Pawlowsky-Glahn, V., and A. Buccianti (2011). Compositional Data Analysis. Theory and Applications. New York: Wiley.

Pawlowsky-Glahn, V., and J.J. Egozcue (2011). Exploring compositional data with the CoDaDendrogram. Austrian Journal of Statistics 40(1\&2), 103-113.

Pawlowsky-Glahn, V., J.J. Egozcue, and D. Lovell (in press). Tools for compositional data with a total. Statistical Modelling. DOI: 10.1177/1471082X14535526.

Ronning, G. (1992). Share equations in econometrics: a story of repression, frustration and dead ends. Statistical Papers 33(1), 307-334.

Thrane, C. (2014). Modelling micro-level tourism expenditure: recommendations on the choice of independent variables, functional form and estimation technique. Tourism Economics 20(1), 51-60.

Thrane, C. (in press). Analysing related choices in tourism research: an intuitive and non-technical primer. Tourism Economics. DOI: 10.5367/te. 2014.0453

Thió-Henestrosa, S., and J.A. Martín-Fernández (2005). Dealing with compositional data: The freeware CoDaPack. Mathematical Geology 37(7), 773-793.

Wang, Y., and M.C.G. Davidson (2010). A review of micro-analyses of tourist expenditure. Current Issues in Tourism 13(6), 507-524

Wang, Y., P. Rompf, D. Severt, and N. Peerapatdit (2006). Examining and identifying the determinants of travel expenditure patterns. International Journal of Tourism Research 8(5), 333-346.

# Changing the reference measure in the simplex and its weighting effects 

J. J. Egozcue ${ }^{1}$, and V. Pawlowsky-Glahn ${ }^{2}$<br>${ }^{1}$ Universitat Poliècnica de Catalunya, Barcelona, Spain; juan.jose.egozcue@upc.edu<br>${ }^{3}$ Universitat de Girona, Girona, Spain


#### Abstract

Standard analysis of compositional data under the assumption that the Aitchison geometry holds assumes a uniform distribution as reference measure of the space. Weighting of parts can be done changing the reference measure. The changes that appear in the algebraic-geometric structure of the simplex are analysed, as a step towards understanding the implications for elementary statistics of random compositions.


Key words: Simplex, sigma-additive measures, subcomposition, weighting, Bayes space

## 1 Introduction

When analysing a composition, some parts may heavily influence the results. For instance, inaccuracies in the measurements in some not fully relevant parts can dominate the analysis, producing a large contribution to variability or to distances. Also, relevance of parts in a given problem can call for weighting techniques to adapt the simplex geometry accordingly. There are a number of weighting techniques that can be useful in this sense. Among them, the change of reference measure of the simplex has several implications that need to be fully understood for a consistent analysis. This contribution is aimed at showing changes that appear in the algebraic-geometric structure of the simplex.

In compositional analysis one of the most fruitful concepts is that of subcomposition (Aitchison, 1986). Some reasonable principles for an analysis of subcompositions to be coherent were established in Aitchison (1992). Beyond the idea that compositional analyses should be scale invariant, they included the assumption that distances between compositions should be greater than or equal to those observed in a subcomposition. This principle, called subcompositional dominance (Aitchison, 1992; Aitchison et al., 2000; Egozcue, 2009), points out a change of the geometry of subcompositions (for instance, a change in inter-distances between two data-points in the subcomposition) with respect to the original geometry of the full composition. Taking a subcomposition can be considered as an extreme case of down-weighting, since the influence of some parts of the composition is removed from the analysis. However, there are cases in which the complete removal of the influence of some parts of the original composition is not desirable. This motivates the idea of weighting compositions as a continuous transition from the full composition, endowed with the corresponding Aitchison geometry (Pawlowsky-Glahn and Egozcue, 2001), to a subcomposition, endowed with the induced Aitchison geometry, which differs in dimension and metrics (distances, inner product, norm).

Apparently, there are many ways of weighting compositions performing a transition from a full composition to a subcomposition. However, fulfilling all coherence requirements is quite challenging. One option deserving attention is the one proposed for Bayes spaces (van den Boogaart et al., 2010; Egozcue et al., 2013) and, more specifically, for Bayes Hilbert spaces (Egozcue et al., 2006; van den Boogaart et al., 2014). In the last reference, it is shown that the simplex, endowed with the Aitchison geometry, is a particular case of Bayes Hilbert space. In the development of Bayes

Hilbert spaces, a reference probability measure is introduced as a parameter regulating the geometry of the measures and densities in the space. This kind of approach provides a way of coherently introducing weighting strategies, both in the simplex and in the analysis of compositional data. The present aim is to start studying the change of reference measure in the simplex, being conscious that there is a long way from the general theory of Bayes Hilbert spaces to applications in compositional data analysis. The main difficulties are interpretative, as usual in compositional data analysis.

The structure of the paper is as follows: Section 2 translates the milestones of Bayes Hilbert spaces into the case of compositions, with special emphasis in the role of the reference measure. Section 3 introduces the clr with respect to an arbitrary reference measure in the simplex, following the definition in van den Boogaart et al. (2014) for general Bayes Hilbert spaces. Section 4 gets into details of metric concepts under a change of the reference measure, such as orthogonality, bases, and balances.

## 2 Change of reference measure for compositions

Consider $D$ categories $c_{1}, c_{2}, \ldots, c_{D}$; they represent a partition of a whole, $\Omega$, also called measurable space. A $D$-part composition $\mathbf{x}=\left(x_{1}, x_{2}, \ldots, x_{D}\right)$ in the $D$-part simplex $\mathcal{S}^{D}$ assigns a proportion $x_{i}$ to the category $c_{i}$. Assuming that the composition is closed to 1 , the proportion assigned to the whole space $\Omega$ is just 1 . For any subset of categories, the proportion assigned is the sum of the corresponding proportions. For instance, the proportion assigned to the subsets $\left\{c_{1}\right\}$ and $\left\{c_{1}, c_{2}\right\}$ are $x_{1}$ and $x_{1}+x_{2}$, respectively. From this point of view, the composition $\mathbf{x}$ defines a finite additive measure on $\Omega$, which is denoted $\mu_{\mathbf{x}}\{\cdot\}$. The argument of this measure is any subset of $\Omega$. Examples are $\mu_{\mathbf{x}}\{\Omega\}=1, \mu_{\mathbf{x}}\{\emptyset\}=0, \mu_{\mathbf{x}}\left\{c_{1}\right\}=x_{1}, \mu_{\mathbf{x}}\left\{c_{1}, c_{2}\right\}=x_{1}+x_{2}$.

Measures can be represented by densities. The idea is that sums (integrals) on a subset of $\Omega$ give the measure of this subset. In the case of the simplex $\mathcal{S}^{D}$, the density is identified with the composition $\mathbf{x}$, as for any subset $A \subseteq \Omega$ it satisfies

$$
\mu_{\mathbf{x}}\{A\}=\sum_{c_{i} \in A} x_{i} P_{0}\left\{c_{i}\right\} \quad, \quad P_{0}\left\{c_{i}\right\}=1, i=1,2, \ldots, D
$$

where the uniform measure $P_{0}\{\cdot\}$ on $\Omega$ has been made explicit as reference measure. Note that $P_{0}\{\Omega\}=D$ and addends of sums (integrals) along the composition are equally weighted with $1=$ $P_{0}\left\{c_{i}\right\}$. The reference measure specified as $\mathbf{p}_{0}=\left(P_{0}\left\{c_{1}\right\}, P_{0}\left\{c_{2}\right\}, \ldots, P_{0}\left\{c_{D}\right\}\right)$ is compositionally equivalent to the neutral element of the simplex. The conclusion is that a composition $\mathbf{x} \in \mathcal{S}^{D}$ defines a measure $\mu_{\mathbf{x}}$ on $\Omega$ specifying the measure of each elementary subset $\left\{c_{i}\right\}$ and, at the same time, $\mathbf{x}$ is the density of $\mu_{\mathbf{x}}$ with respect to the uniform reference measure $P_{0}$, which density is $\mathbf{p}_{0}$. When $P_{0}$ is the uniform reference measure there is no need to distinguish between $\mathbf{x}$ as a composition, as a measure or as a density. These facts change when weights are introduced through the reference measure.

A first step for introducing weights is a change of reference measure. Let be $\mathbf{p}=\left(p_{1}, p_{2}, \ldots, p_{D}\right)$ an arbitrary array of positive weights, as in the case of the uniform reference measure $\mathbf{p}_{0}$. The corresponding measure $P$ is then characterised by $P\left\{c_{i}\right\}=p_{i}$, for $i=1,2, \ldots, D$, and by the measure of the whole space, $P(\Omega)=\sum_{i=1}^{D} p_{i}$. A question is now to look for a composition $\mathbf{y}$ which gives the density of the measure $\mu_{\mathbf{x}}$ with respect to the new reference measure $P$. Note that $\mathbf{p}$ is its density with respect to the uniform measure $P_{0}$. Such a composition is $\mathbf{y}=\mathbf{x} \ominus \mathbf{p}$, where $\mathbf{p}$ is now treated as a composition, despite the fact that $\mathbf{p}$ is not closed to 1 but to $P(\Omega)$. In fact, for $A \subseteq \Omega$,

$$
\mu_{\mathbf{x}}\{A\}=\sum_{c_{i} \in A} y_{i} p_{i}=\sum_{c_{i} \in A} \frac{x_{i}}{p_{i}} p_{i}
$$

thus the measure $\mu_{\mathbf{x}}$ is now recovered from two different densities, $\mathbf{x}$ when considering the uniform reference $P_{0}$ and $\mathbf{y}$ for a reference $P$. If the reference measure is represented by the array of weights
$\mathbf{p}$, the new density is just a perturbation of $\mathbf{x}$, a shift in the simplex $\mathbf{y}=\mathbf{x} \ominus \mathbf{p}$, recalling that the perturbation-difference $\ominus$ includes the closure, $\mathcal{C}$, and consequently $\mathbf{x} \ominus \mathbf{p}=\mathbf{x} \ominus \mathcal{C} \mathbf{p}$. From now on, the composition $\mathbf{x} \ominus \mathbf{p}$ is denoted $\mathbf{y}^{(\mathbf{p})}$, or simply $\mathbf{y}$ if clear from the context.

Perturbation and powering in the simplex change when the reference measure is changed. Specifically, consider $\mathbf{x}_{1}$ and $\mathbf{x}_{2}$ compositions (densities) in $\mathcal{S}^{D}$ with respect to the uniform reference $P_{0}$. Let $\mathbf{y}_{1}^{(\mathbf{p})}, \mathbf{y}_{2}^{(\mathbf{p})}$ be the compositions (densities) when the reference measure is changed to $P$, which weights are in $\mathbf{p}$. The perturbation, $\oplus_{\mathbf{p}}$, and powering, $\odot_{\mathbf{p}}$, in $\mathcal{S}^{D}$ with the reference $P$ operate as

$$
\begin{gathered}
\mathbf{y}_{1}^{(\mathbf{p})} \oplus_{\mathbf{p}} \mathbf{y}_{2}^{(\mathbf{p})}=\mathbf{y}_{1}^{(\mathbf{p})} \oplus \mathbf{y}_{2}^{(\mathbf{p})} \oplus \mathbf{p}=\left[\left(\mathbf{y}_{1}^{(\mathbf{p})} \oplus \mathbf{p}\right) \oplus\left(\mathbf{y}_{2}^{(\mathbf{p})} \oplus \mathbf{p}\right)\right] \ominus \mathbf{p} \\
\alpha \odot_{\mathbf{p}} \mathbf{y}_{1}^{(\mathbf{p})}=\left(\alpha \odot \mathbf{y}_{1}^{(\mathbf{p})}\right) \oplus \mathbf{p}
\end{gathered}
$$

where $\mathbf{y}_{i}^{(\mathbf{p})} \oplus \mathbf{p}=\mathbf{x}_{i}$. Following van den Boogaart et al. (2010) and van den Boogaart et al. (2014), a weighted perturbation and powering can be defined for the compositions $\mathbf{y}_{1}^{(\mathbf{p})}$ and $\mathbf{y}_{2}^{(\mathbf{p})}$ such that they operate linearly in the weighted simplex. However, standard perturbation and powering are easily interpreted and computed in the applications; therefore, its use is recommended, thus avoiding linear operations with the shifted densities $\mathbf{y}^{(\mathbf{p})}=\mathbf{x} \ominus \mathbf{p}$. In practice, the weighted compositions will be used in the computation of distances and inner products as explained in the next Sections.

## 3 Centred log-ratio with respect to a reference measure

In van den Boogaart et al. (2014), the clr-transformation of a density $f$ with respect to a given reference measure $P$, is defined as

$$
\begin{equation*}
\operatorname{clr}_{P}(f)(x)=\log f(x)-\frac{1}{P\{\Omega\}} \int_{\Omega} \log f(\xi) d P(\xi), \quad x \in \Omega \tag{1}
\end{equation*}
$$

where $\Omega$ is the measurable set where the density $f$ is defined. In the present case $\Omega$ is the set of the $D$ parts or categories of $\mathcal{S}^{D}$, named $c_{i}, i=1,2, \ldots, D$. Therefore, the values of $x$ in such an expression correspond to the $c_{i}$ 's. Since $f$ is a density of a measure with respect to the reference measure $P$, it can be identified with the composition $\mathbf{y}=\mathbf{x} \ominus \mathbf{p}$ as introduced in Section 2. With these identifications, the clr $_{\mathbf{p}}$-transformation of the simplex with respect to the measure $P$, represented by $\mathbf{p}$, is

$$
\begin{equation*}
\operatorname{clr}_{\mathbf{p}}(\mathbf{y})=\left(\log \frac{y_{1}}{\mathrm{~g}_{\mathbf{p}}(\mathbf{y})}, \log \frac{y_{2}}{\mathrm{~g}_{\mathbf{p}}(\mathbf{y})}, \ldots, \log \frac{y_{D}}{\mathrm{~g}_{\mathbf{p}}(\mathbf{y})}\right), \quad \mathrm{g}_{\mathbf{p}}(\mathbf{y})=\exp \left(\frac{1}{s_{\mathbf{p}}} \sum_{i=1}^{D} p_{i} \log y_{i}\right) \tag{2}
\end{equation*}
$$

where $s_{\mathbf{p}}=\sum_{i=1}^{D} p_{i}$, and $\mathrm{g}_{\mathbf{p}}(\cdot)$ denotes a weighted geometric mean of the parts $y_{i}$. Note that $\mathbf{p}$, the reference measure of the categories $c_{i}$, is not closed to $D$, and $P\{\Omega\}=s_{\mathbf{p}}$. The fact that, for $P_{0}$ as uniform reference measure, $s_{\mathbf{p}_{0}}=D$ is also remarkable. An important characteristic of $\operatorname{clr}_{\mathbf{p}}(\mathbf{y})$ is that the weighted sum of its $D$ components is zero, that is

$$
\begin{equation*}
\sum_{i=1}^{D} p_{i} \log \frac{y_{i}}{\mathrm{~g}_{\mathbf{p}}(\mathbf{y})}=0 \tag{3}
\end{equation*}
$$

generalising the ordinary clr in $\mathcal{S}^{D}$, for which the sum of its components (weights equal to 1 ) is null. This has a geometric interpretation in the space $\mathbb{R}^{D}$, where a point has coordinates $\log (\mathbf{y})=\left(\log y_{1}, \log y_{2}, \ldots, \log y_{D}\right)$. As illustrated in Figure 1, which shows a scheme for $D=2$, to obtain the ordinary clr of a generic point $\log (\mathbf{y})$, the point is orthogonally projected onto a hyperplane through the origin whose orthogonal vector is $(1,1, \ldots, 1)$ (Aitchison, 1986; PawlowskyGlahn et al., 2015). When using a non-neutral $\mathbf{p}=\left(p_{1}, p_{2}, \ldots, p_{D}\right)$ the procedure to $\operatorname{get}^{\operatorname{clr}}(\mathbf{p}(\mathbf{y})$ is


Figure 1: Generic 2-part composition $\left(x_{1}, x_{2}\right)$, $\log$-transformed into $\left(\log x_{1}, \log x_{2}\right)$ in $\mathbb{R}_{+}^{2}$. Two reference measures with densities $(1,1)$ (uniform, blue arrow) and $\mathbf{p}$ (red arrow) are considered. The point $\left(\log x_{1}, \log x_{2}\right)$ is projected, parallel to the reference arrow, on the clr-plane (blue) and on the $\operatorname{clr}_{\mathbf{p}}$-plane (red), thus obtaining the respective transformations.
to orthogonally project the point $\log (\mathbf{y})$ onto a hyperplane whose orthogonal vector is $\mathbf{p}$, as shown by the inner product in $\mathbb{R}^{D}$ implicit in Equation (3). Summarising, clr $\mathbf{p}_{\mathbf{p}}$ is a projection of $\log (\mathbf{y})$ on a hyperplane whose normal vector is $\mathbf{p}$.

A particular case of interest is that of

$$
\begin{equation*}
p_{i}=1, i=1,2, \ldots, D-1, p_{D}=\epsilon \tag{4}
\end{equation*}
$$

for which $P(\Omega)=(D-1)+\epsilon$. When $\epsilon \rightarrow 0$, the $D$-th part is down-weighted from 1 to $\epsilon \ll 1$. For small enough $\epsilon$, the weighted geometric mean $\mathrm{g}_{\mathbf{p}}$ in Equation (2) approaches the ordinary geometric mean of the first $D-1$ parts of $\mathbf{y}$. A consequence is that the first $D-1$ components of $\operatorname{clr}_{\mathbf{p}}(\mathbf{y})$ approach the ordinary clr of the subcomposition formed by $\left(y_{1}, y_{2}, \ldots, y_{D-1}\right)$. This suggests that this kind of reference measures may approach the induced Aitchison geometry on the subcomposition.

## 4 Metrics under change of reference

The clr transformation can be used to define the inner product in $\mathcal{S}^{D}$, as was done in Bayes Hilbert spaces (van den Boogaart et al., 2014). There, the proposed definition was

$$
\left\langle\mathbf{y}_{1}, \mathbf{y}_{2}\right\rangle_{B^{2}}=\frac{1}{P\{\Omega\}}\left\langle\operatorname{clr}_{P}\left(\mathbf{y}_{2}\right), \operatorname{clr}_{P}\left(\mathbf{y}_{1}\right)\right\rangle
$$

where $\langle\cdot, \cdot\rangle$ is the ordinary inner product in $\mathbb{R}^{D}$. This definition leads to an inner product in $\mathcal{S}^{D}$ which, for a uniform reference measure $\mathbf{p}=(1,1, \ldots, 1)$, is

$$
\begin{equation*}
\left\langle\mathbf{y}_{1}, \mathbf{y}_{2}\right\rangle_{a}=\frac{1}{D}\left\langle\operatorname{clr}\left(\mathbf{y}_{2}\right), \operatorname{clr}\left(\mathbf{y}_{1}\right)\right\rangle \tag{5}
\end{equation*}
$$

which is not the standard in compositional data analysis due to the factor $1 / D$. This inner product is not suitable for compositional data analysis, as it does not fulfill the principle of subcompositional dominance of distances. For instance, consider the 3 -part compositions $\mathbf{u}=(0.1,0.7,0.2)$ and $\mathbf{v}=(1 / 3,1 / 3,1 / 3)$. Their distance, in the geometry induced by the inner product (5) in $\mathcal{S}^{3}$, is $\mathrm{d}_{3}(\mathbf{u}, \mathbf{v})=0.805$. Taking the subcomposition formed by the first and second part and computing the distance in $\mathcal{S}^{2}$ according to (5), the result is $\mathrm{d}_{2}(\mathbf{u}, \mathbf{v})=0.973$. Since $\mathrm{d}_{3}(\mathbf{u}, \mathbf{v})<\mathrm{d}_{2}(\mathbf{u}, \mathbf{v})$, the principle of subcompositional dominance is violated.

To match the present definition to the standard practice in compositional data analysis (Aitchison, 1986; Aitchison and Egozcue, 2005; Egozcue et al., 2011; Pawlowsky-Glahn et al., 2015) and the subcompositional dominance of distances, the factor $1 / D$ in (5) is suppressed.

In the case of using a reference measure represented by the weights $\mathbf{p}$, the appropriate definition of the weighted Aitchison inner product is

$$
\begin{equation*}
\langle\mathbf{y}, \mathbf{z}\rangle_{\mathbf{p}}=\sum_{i=1}^{D} p_{i} \log \frac{y_{i}}{\mathrm{~g}_{\mathbf{p}}(\mathbf{y})} \log \frac{z_{i}}{\mathrm{~g}_{\mathbf{p}}(\mathbf{z})} \tag{6}
\end{equation*}
$$

where $\mathbf{y}=\mathbf{y}^{(\mathbf{p})}, \mathbf{z}=\mathbf{z}^{(\mathbf{p})}$ are in $\mathcal{S}^{D}$. The expression in the right hand side of Equation (6) is an inner product of the involved $\operatorname{clr}_{\mathbf{p}}$ as real vectors with respect to the measure $P$. Note that, for the reference measures in Equation (4), the weights of the reference measure add neither to one nor to $D$.

The weighted Aitchison norm is readily derived from the inner product as $\|\mathbf{y}\|_{\mathbf{p}}^{2}=\langle\mathbf{y}, \mathbf{y}\rangle_{\mathbf{p}}$. Also the distance comes from the inner product. However, it is preferable to give the explicit expression

$$
\mathrm{d}_{\mathbf{p}}^{2}(\mathbf{y}, \mathbf{z})=\sum_{i=1}^{D} p_{i}\left(\log \frac{y_{i}}{\mathrm{~g}_{\mathbf{p}}(\mathbf{y})}-\log \frac{z_{i}}{\mathrm{~g}_{\mathbf{p}}(\mathbf{z})}\right)^{2}
$$

These definitions coincide with those of the ordinary Aitchison geometry of $\mathcal{S}^{D}$ whenever $\mathbf{p}=$ $(1,1, \ldots, 1)$. When the weights $\mathbf{p}$ are not uniform, the inner product differs from the ordinary Aitchison inner product and, consequently, also norm and distance are different.

To get a further intuition of what is changing with $\mathbf{p}$, it is instructive to build orthonormal basis of the simplex according to the change of reference. It allows to show how these bases are under change of $\mathbf{p}$ in particular cases.

A straightforward technique for obtaining orthonormal basis of the simplex and their respective coordinates is that of sequential binary partitions (SBP) (Egozcue et al., 2003; Egozcue and Pawlowsky-Glahn, 2005, 2006). When using a non-uniform reference measure with the weights $\mathbf{p}$, the procedure is based on partition coded as in Table 1, but the formulae to obtain the contrast matrix are modified. Table 1 shows a generic sign code for an SBP, adding weights $\mathbf{p}$ as column labels (second row) for further comment on the generalised technique.

Table 1: A generic table of an SBP for a five-part composition. Weights from the reference measure are placed in the second row, under the part label. Rows are labelled as balances $b_{i}$ for further reference.

| parts | $y_{1}$ | $y_{2}$ | $y_{3}$ | $y_{4}$ | $y_{5}$ |
| :--- | ---: | ---: | ---: | ---: | ---: |
| weights | $p_{1}$ | $p_{2}$ | $p_{3}$ | $p_{4}$ | $p_{5}$ |
| $b_{1}$ | +1 | -1 | -1 | -1 | +1 |
| $b_{2}$ | +1 | 0 | 0 | 0 | -1 |
| $b_{3}$ | 0 | +1 | -1 | -1 | 0 |
| $b_{4}$ | 0 | 0 | +1 | -1 | 0 |

Denote the entries of the matrix code as $\theta_{i j}, i=1,2, \ldots, D-1, j=1,2, \ldots, D$. For the case in Table $1, D=5$ and, for instance, $\theta_{32}=+1$. When using the standard uniform reference measure with $\mathbf{p}=(1,1, \ldots, 1)$, the clr coefficients of a balancing element in the basis are given by

$$
\psi_{i j}= \begin{cases}+\frac{1}{n_{i}^{+}} \sqrt{\frac{n_{i}^{+} n_{i}^{-}}{n_{i}^{+}+n_{i}^{-}}} & \text {if } \theta_{i j}=+1  \tag{7}\\ -\frac{1}{n_{i}^{-}} \sqrt{\frac{n_{i}^{+} n_{i}^{-}}{n_{i}^{+}+n_{i}^{-}}} & \text {if } \theta_{i j}=-1 \\ 0 & \text { if } \theta_{i j}=0\end{cases}
$$

where $n_{i}^{+}, n_{i}^{-}$denote the number of $+1,-1$ in the $i$-th row of the sign code table, respectively.
When using the reference measure which weights $p_{j}$ are not uniform, these formulas for the $\operatorname{clr}_{\mathbf{p}}$ of balancing elements are the same except that $n_{i}^{+}, n_{i}^{-}$are

$$
n_{i}^{+}=\sum_{\theta_{i j}=+1} p_{j} \quad, \quad n_{i}^{-}=\sum_{\theta_{i j}=-1} p_{j} .
$$

The contrast matrix $\Psi$, with entries $\psi_{i j}, i=1,2, \ldots, D, j=1,2, \ldots, D-1$, fulfils the conditions

$$
\begin{equation*}
\Psi \operatorname{diag}(\mathbf{p}) \Psi^{\top}=I_{D-1} \quad, \quad \operatorname{diag}(\mathbf{p}) \Psi^{\top} \Psi=I_{D}-\frac{1}{D} \mathbf{p}^{\top} \mathbf{1} \tag{8}
\end{equation*}
$$

where $I_{m}$ is the $(m, m)$-identity matrix; $\mathbf{p}$ and $\mathbf{1}=(1,1, \ldots, 1)$ are taken as row vectors, and $\operatorname{diag}(\mathbf{p})$ is a $(D, D)$ diagonal matrix with entries equal to the components of $\mathbf{p}$. The first condition is equivalent to saying that balancing elements are unitary compositions mutually orthogonal. In fact, their $\operatorname{clr}_{\mathbf{p}}$ are unitary and orthogonal in the weighted Euclidean geometry.

Coordinates of a composition $\mathbf{y} \in \mathcal{S}^{D}$ with respect to an orthonormal basis are found carrying out the inner product of a balancing element in the basis with the composition $\mathbf{y}=\mathbf{x} \ominus \mathbf{p}$. These coordinates will be named weighted isometric log-ratio coordinates and denoted by ilr $\mathbf{p}$. In the particular case in which they are obtained using an SBP, they are called weighted balances. For simplicity, these weighted balances are denoted $b_{i}, i=1,2, \ldots, D-1$, with no reference to the weights associated with the change of measure (as shown in Table 1). The ilr $\mathbf{p}_{\mathbf{p}}$ coordinates can be obtained using the matrix expression

$$
\begin{equation*}
\operatorname{ilr}_{\mathbf{p}}(\mathbf{y})=\mathbf{b}=\operatorname{clr}_{\mathbf{p}}(\mathbf{y}) \operatorname{diag}(\mathbf{p}) \Psi^{\top} \tag{9}
\end{equation*}
$$

where compositions and their $\operatorname{clr}_{\mathbf{p}}$, ilr $_{\mathbf{p}}$ transforms are considered row-vectors. Note that each component of $\mathbf{b}=\left(b_{1}, b_{2}, \ldots, b_{D}\right)$ is a weighted inner product of $\operatorname{clr}_{\mathbf{p}}(\mathbf{y})$ with the corresponding $\operatorname{clr}_{\mathbf{p}}$ of a balancing element. The inverse ilr $\mathbf{p}_{\mathbf{p}}$ transformation is readily obtained using the properties (8) of $\Psi$

$$
\mathbf{y}=\mathcal{C} \exp \left(\Psi^{\top} \operatorname{ilr}_{\mathbf{p}}(\mathbf{y})\right),
$$

which is formally identical to the inverse ilr with uniform reference measure.
Although Equation (9) is useful from a computational point of view, an explicit expression of balances gives a more detailed insight on the meaning of weighted balances. Consider a sign code of a step in an SBP, for which $n_{i}^{+}, n_{i}^{-}$are given. The corresponding weighted balance is

$$
b_{i}=\sqrt{\frac{n_{i}^{+} n_{i}^{-}}{n_{i}^{+}+n_{i}^{-}}} \log \left(\frac{\left(\prod_{\left(\theta_{i j}=+1\right)} y_{j}^{p_{j} / n_{i}^{+}}\right)}{\left(\prod_{\left(\theta_{i j}=-1\right)} y_{j}^{p_{j} / n_{i}^{-}}\right)}\right)
$$

where the products span on the parts corresponding to the sign code $\theta_{i j}$. When the weights $p_{j}=1$, the balance reduces to the standard balances, as $n_{i}^{+}, n_{i}^{-}$are then the number of +1 and -1 in the $i$-th row of the sign code respectively. The main feature, when the reference is not uniform, is that the ratios within the logarithm are ratios of a kind of weighted geometric means. Note that, in general, $n_{i}^{+}, n_{i}^{-}$are not integers and each part is powered to the weight corresponding to that part. When some $p_{j}$ is small, relative to other weights, it plays a minor role in these weighted geometric means. Furthermore, the weighted balances are scale invariant log-contrasts, that is, if the composition $\mathbf{y}$ is multiplied by a positive constant, the weighted balance remains unaltered.

A critical point whenever there is a change in the geometry of compositions is the subcompositional dominance of distances. In the standard approach, the distance between any two compositions $\mathbf{x}_{1}$, $\mathbf{x}_{2} \in \mathcal{S}^{D}$ is $\mathrm{d}_{a}\left(\mathbf{x}_{1}, \mathbf{x}_{2}\right)$. After taking a given subcomposition in $\mathcal{S}^{d}, d<D$, the distance between the respective subcompositions, $\mathbf{x}_{1}^{(d)}, \mathbf{x}_{2}^{(d)}$, satisfies $\mathrm{d}_{a}\left(\mathbf{x}_{1}^{(d)}, \mathbf{x}_{2}^{(d)}\right) \leq \mathrm{d}_{a}\left(\mathbf{x}_{1}, \mathbf{x}_{2}\right)$. In this case, both spaces have integer reference measures $P\left(\Omega_{D}\right)=D$ and $P\left(\Omega_{d}\right)=d$ and the corresponding weights


Figure 2: Evolution of weighted square-distances between three measures represented by the compositions $\mathbf{x}_{1}=$ $(0.1,0.7,0.2), \mathbf{x}_{2}=(0.5,0.3,0.2), \mathbf{x}_{3}=(0.9,0.08,0.02)$ with respect to the uniform reference in $\mathcal{S}^{3}$. With $\mathbf{y}_{i}=\mathbf{x} \ominus \mathbf{p}$, square-distance curves are $\mathrm{d}_{\mathbf{p}}\left(\mathbf{y}_{1}, \mathbf{y}_{2}\right)$ (black), $\mathrm{d}_{\mathbf{p}}\left(\mathbf{y}_{1}, \mathbf{y}_{3}\right)$ (blue), $\mathrm{d}_{\mathbf{p}}\left(\mathbf{y}_{2}, \mathbf{y}_{3}\right)$ (red). Reference measure is $\mathbf{p}=(1,1, \epsilon)$ and x-axis is scaled as $P(\Omega)=1+1+\epsilon$. The three square-distances monotonically increase from $P(\Omega)=2$ to $P(\Omega)=3$. The end points of the curves at $P(\Omega)=2$ and $P(\Omega)=3$ are equal to standard Aitchison square-distances in $\mathcal{S}^{2}$ and $\mathcal{S}^{3}$ respectively.


Figure 3: The unit circle (black) in the uniform reference. After change of origin to $(1,1, \epsilon), \epsilon=0.5$ (brown), 0.1 (blue), 0.05 (green), and 0.01 (red), the circle is shifted towards the vertex $y_{3}$
are $(1,1,0)$ and $(1,1,1)$ respectively. When changing reference measure by down weighting some of the weights, it is expected a dominance of distances as it occurs when taking subcompositions. The dominance of distances can be stated as follows.
Proposition (dominance of distances) Let $\mathbf{x}_{1}, \mathbf{x}_{2}$ be two compositions in $\mathcal{S}^{D}$, endowed with the uniform reference measure, which weights are $(1,1, \ldots, 1)$. Consider two reference measures, $P_{1}$ and $P_{2}$, represented by their respective weights $\mathbf{p}_{1}=\left(p_{11}, p_{12}, \ldots, p_{1 D}\right)$ and $\mathbf{p}_{2}=\left(p_{21}, p_{22}, \ldots, p_{2 D}\right)$, such that all their components are $0<p_{k i} \leq 1$, for $k=1,2, i=1,2, \ldots, D$ and $P_{k}(\Omega)=\sum_{i=1}^{D} p_{k i}$. Define the compositions $\mathbf{y}_{j}^{\left(\mathbf{p}_{k}\right)}=\mathbf{x}_{j} \ominus \mathbf{p}_{k}$ for $k=1,2$ and $j=1,2$. Then,

$$
p_{1 i} \leq p_{2 i}, i=1,2, \ldots, D \quad \text { implies } \quad \mathrm{d}_{\mathbf{p}_{1}}\left(\mathbf{y}_{1}^{\left(\mathbf{p}_{1}\right)}, \mathbf{y}_{2}^{\left(\mathbf{p}_{1}\right)}\right) \leq \mathrm{d}_{\mathbf{p}_{2}}\left(\mathbf{y}_{1}^{\left(\mathbf{p}_{2}\right)}, \mathbf{y}_{2}^{\left(\mathbf{p}_{2}\right)}\right)
$$

It is worth to remark that the notation of distances like $d_{\mathbf{p}_{1}}\left(\mathbf{y}_{1}^{\left(\mathbf{p}_{1}\right)}, \mathbf{y}_{2}^{\left(\mathbf{p}_{1}\right)}\right)$ could be changed to $\mathrm{d}_{\mathbf{p}_{1}}\left(\mathbf{x}_{1}, \mathbf{x}_{2}\right)$, as distances assigned to shifted $\mathbf{y}$ 's are taken as equal to original compositions $\mathbf{x}^{\prime}$ 's. This is due to the fact that $\mathbf{x}$ and $\mathbf{y}$ are densities of the same measure, with respect to different reference measures.

Figure 2 shows the evolution of square-distances between three compositions $\mathbf{x}_{1}=(0.1,0.7,0.2)$, $\mathbf{x}_{2}=(0.5,0.3,0.2), \mathbf{x}_{3}=(0.9,0.08,0.02)$ with respect to the uniform reference in $\mathcal{S}^{3}$ when the reference measure changes progressively. The reference measure is $(1,1, \epsilon)$, with epsilon going from 0 to 1 . The plot is scaled according the $P(\Omega)=1+1+\epsilon$. The square-distances increase monotonically, from distances corresponding to the subcomposition $\left(y_{1}, y_{2}\right)$ to square-distances with the standard reference $\mathbf{p}_{0}=(1,1,1)$. This result is expected after the previous proposition.

An experiment has been conducted to show how the changes of reference modify distances and shapes. Five different reference measures $\mathbf{p}=(1,1, \epsilon)$ have been considered with $\epsilon$ equal to $1,0.5$, $0.1,0.05,0.01$, so that they approach progressively the geometry of the subcomposition of the two first parts. The unit circle centered at the neutral element was shifted by the five reference measures. Figure 3, shows this unit circle (black) and the sequence of perturbations as a consequence of the change of origin. Note that the transformed circle is shifted to the vertex which weight is reduced, as expected after dividing each part by the corresponding weight.

After the change of origin, each point on the circles was ilr-transformed using the corresponding weights according the SBP sign code

| $y_{1}$ | $y_{2}$ | $y_{3}$ |
| ---: | ---: | ---: |
| +1 | -1 | -1 |
| 0 | +1 | -1 |

which has been selected to avoid a balance representing the subcomposition $y_{1}, y_{2}$. Figure 4 (left panel) shows the coordinates of the circles to show the changes of the distances between points on the same circle. Note that the centers of the ellipses do not coincide, as they correspond to the closure of the reference measure $(1,1, \epsilon)$. The main feature is the progressive stretch of the original circle. For very small $\epsilon$ the ellipse tends to degenerate into a segment following the direction of the subcomposition $y_{1}, y_{2}$. Similarly, Figure 4 (right panel) shows the deformation of a grid originally at $-1,0,1$ in both axes (black). The new references are $\epsilon=0.1$ (blue), and 0.01 (red). The grid is progressively tilted and distances between nodes decrease as $\epsilon$ also decreases. Although straightlines are preserved, their angles change, thus showing the change of geometry when changing the reference.

## 5 Conclusions and further research

A weighting strategy for the analysis of compositions is proposed. It is based on the theory of Bayes Hilbert spaces. However, some modifications have been introduced to fulfill the desirable principle of dominance of distances when down-weighting some parts of the composition. When


Figure 4: Left panel: the five circles in Figure 3 after weighted ilr-transformation. Reference measures are $(1,1, \epsilon)$, $\epsilon=1$ (black), $\epsilon=0.5$ (brown), 0.1 (blue), 0.05 (green), and 0.01 (red). Left panel: a regular grid at points -1 , $0,-1$ in both axes after change of origin and weighted ilr transformation. Weights are $(1,1, \epsilon), \epsilon=1$ (black), 0.1 (blue), and 0.01 (red).
the weights considered are unitary in each part, that is, when there is no down-weighting, the approach is reduced to the standard compositional data analysis. If some parts are down-weighted approaching zero, the weighted geometry of the simplex reduces to the ordinary Aitchison geometry of the corresponding subcomposition.

In order to use the proposed weighting approach, it seems advisable to deal with compositions as usual for linear operations, using the standard perturbation and powering. When distances or inner products are involved in the analysis, they are computed in two steps: first, shifting the compositions by $\ominus \mathbf{p}$, that is, dividing each part by the corresponding weight; and second, computing $\operatorname{clr}_{\mathbf{p}}$ or $\operatorname{ilr}_{\mathbf{p}}$ to find the required distances or inner products in a straightforward way.

Statistical consequences of weighting compositions need to be studied in the future. Standard tools of exploratory analysis, as variation matrix, biplots or balance dendrograms, clustering and others, will be affected by weighting. The reason is that distances between compositions and computation of variances-covariances are affected as well. As a consequence the proposed weighting approach is only a first step towards developing effective weighting techniques applicable to compositional data analysis.

## Acknowledgements

This research has been supported by the Spanish Ministry of Education and Science under project 'METRICS' (Ref. MTM2012-33236); and by the Agència de Gestió d’Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project Ref. 2009SGR424.

## References

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. London (UK): Chapman \& Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Aitchison, J. (1992). On Criteria for Measures of Compositional Difference. Mathematical Geology $24(4), 365-379$.

Aitchison, J., C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2000). Logratio analysis and compositional distance. Mathematical Geology 32(3), 271-275.

Aitchison, J. and J. J. Egozcue (2005). Compositional data analysis: where are we and where should we be heading? Mathematical Geology 37(7), 829-850.

Egozcue, J. J. (2009). Reply to "On the Harker variation diagrams;..." by J. A. Cortés. Mathematical Geosciences 41 (7), 829-834.

Egozcue, J. J., C. Barceló-Vidal, J. A. Martín-Fernández, E. Jarauta-Bragulat, J.-L. DíazBarrero, and G. Mateu-Figueras (2011). Elements of simplicial linear algebra and geometry. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), Compositional Data Analysis: Theory and Applications., pp. 141-157. John Wiley \& Sons. 378 p.

Egozcue, J. J., J. L. Díaz-Barrero, and V. Pawlowsky-Glahn (2006). Hilbert space of probability density functions based on Aitchison geometry. Volume 22, pp. 1175-1182. DOI: 10.1007/s10114-005-0678-2.

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

Egozcue, J. J. and V. Pawlowsky-Glahn (2006). Simplicial geometry for compositional data. In Compositional Data Analysis in the Geosciences: From Theory to Practice, pp. 145-159.

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

Egozcue, J. J., V. Pawlowsky-Glahn, R. Tolosana-Delgado, M. I. Ortego, and K. G. van den Boogaart (2013). Bayes spaces: use of improper distributions and exponential families. Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales, Serie A, Matemáticas (RACSAM) 107, 475-486. DOI 10.1007/s13398-012-0082-6.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment (SERRA) 15(5), 384-398.

Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2015). Modeling and analysis of compositional data. Statistics in practice. John Wiley \& Sons, Chichester UK. 272 pp.
van den Boogaart, K., J. Egozcue, and V. Pawlowsky-Glahn (2010). Bayes linear spaces. SORT Statistics and Operations Research Transactions $34(2), 201-222$.
van den Boogaart, K. G., J. J. Egozcue, and V. Pawlowsky-Glahn (2014). Bayes Hilbert spaces. Australian and New Zealand Journal of Statistics 56(2), 171-194.

# Semi-parametric density estimation with TrioScale 

Paul H. C. Eilers ${ }^{1}$, Mark de Rooij ${ }^{2}$<br>${ }^{1}$ Erasmus University Medical Center, Rotterdam, the Netherlands; p.eilers@erasmusmc.nl<br>${ }^{2}$ Leiden University, Leiden, the Netherlands


#### Abstract

The TrioScale display transforms three-dimensional compositions to unbounded twodimensional coordinates. Then it is easy to estimate a smooth density for better visualization of scatterplots. We use penalized least squares smoothing of a detailed twodimensional histogram and back-transformation for display of a density in a ternary diagram. Originally we derived TrioScale by representing logarithms of proportions as negative squared distances from anchor points. An equivalent but simpler interpretation uses cyclic log-ratios, resulting in an unbounded triangular diagram with linear log-ratio axes. We also establish an equivalence to the isometric log-ratio transform. We apply our smoothing procedure to real-world data.


Key words: Smoothing, compositional data, visualization.

## 1 Introduction

Anyone doing serious work on compositional data is familiar with the ternary diagram. It is an insightful way of presenting observations, but it also has limitations. The most obvious one is that all data points have to lie within an isosceles triangle, or on its boundary. Hence small fractions are hard to judge, the same is true for fractions close to 1 .

When the number of data points is large, points start to overlap each other and patterns become hard to judge. This problem is not unique to ternary diagrams. Any scatterplot with many dots suffers from it. A solution is to estimate the local density of the data points and color code it, to get an intensity image. Apart from making the picture more attractive and more clear, it also makes graphic files smaller, reducing demands on storage, transmission speed and printing time.

Several approaches have been proposed and most of them have been implemented in R software. The hexbin plot, in the R package hexbin (Carr, 2014), divides the $x-y$ plane into small hexagons, counts the number of observations in each of them and plots them, filled with colors proportional to the counts. The scatterplot smoother of Eilers and Goeman (2004) counts observations in small bins on a rectangular grid, and applies penalized least squares smoothing in both directions of the grid. Kernel smoothers replace each observation by a small "hill" with a two-dimensional Gaussian shape and adds up all the hills, to get a smooth surface that can be evaluated at any desired resolution. In that way a color-coded image of the density can be obtained and displayed.

When we consider using these smoothing algorithms to compositional data, we encounter various problems. The scatterplot smoother needs a complete rectangular grid, so it cannot be applied directly. The hexbin smoother might work, if we cover the triangle with small enough hexagons. As far as we are aware, no implementation is available. The kernel smoother has been implemented in the ggtern package. It presents attractive figures, with color gradients and level contours. But it also suffers from the typical kernel smoother disease: spreading beyond the boundaries of the data. An example is shown in Figure 1. The cloud of data points gets quite narrow in the right corner, but this is not reflected in the density estimate. It has to be cut off sharply at the edges of the triangle.

To get out of his rut, we propose to first transform the data to another, unbounded, domain, and use the scatterplot smoother there. We use the TrioScale transformation, which was proposed by us at CoDaWork 2013 (de Rooij and Eilers, 2013). We believe it to be an attractive way to represent compositional data and the results of (logistic) models. If one shares our conviction, the final result would be the smooth density in TrioScale coordinates. Nevertheless, it is useful and interesting to transform back to the ternary diagram. To do this properly, one has to apply the proper geometrical back-transformation to match locations in both coordinate systems, but also correct the density values by the Jacobian of the transformation.

The plan of the paper is as follows. To make it self-contained, we present a summary of TrioScale in Section 2. There we also show that, although TrioScale is defined by distances to anchor points in the plane of new coordinates, it can also be expressed in logs of cyclic ratios of fractions $\left(p_{1} / p_{2}\right.$, $p_{2} / p_{3}$ and $\left.p_{3} / p_{1}\right)$. In Section 3 we discuss the fitting and display of bivariate normal distributions. In section 4 presents the scatterplot smoother and applies it to simulated and real data. The backtransformation to the ternary diagram is discussed there too. Section 5, the Discussion, closes the paper.


Figure 1: An example of kernel density smoothing in the ternary diagram. Left: data points. Right density estimate as produced by the R package ggtern.

## 2 TrioScale in a nutshell

The origins of TrioScale lie in the ideal point classification model (De Rooij, 2009) In the cases of three classes there are three class points (to be estimated from the data) with coordinates ( $u_{j}, v_{j}$ ), for $j=1, \ldots$, A triple of proportions, $p$, is represented by a point with coordinates $(u, v)$ in such a way that $p_{j}=c \exp \left(-d_{j}^{2}\right)$, where $d_{j}^{2}=\left(u-u_{j}\right)^{2}+\left(v-v_{j}\right)^{2}$, the squared distance to the corresponding class point. The value of $c$ is chosen such that $\sum_{j} p_{j}=1$. In the Appendix we show how to compute $u$ and $v$ for given $p$ and class point positions. However, we will call them anchor points, because in TrioScale we do not estimate them, but set their positions ourselves, usually forming a triangle with equal sides, like in the familiar ternary diagram.

An alternative interpretation is based on cyclic log-ratios. Figure 2 shows three anchor points and a line perpendicular to the line connecting $A$ and $B$ and crossing it at $D$. It is easy to prove that $|A E|^{2}-|B E|^{2}=|A D|^{2}-|B D|^{2}$, independent of the position of point $E$ on the perpendicular. If $A$ has coordinates $\left(u_{1}, v_{1}\right), B$ coordinates $\left(u_{2}, v_{2}\right)$ and $E$ coordinates $(u, v)$, then we have that $\log p_{1}-\log p_{2}=b$, a constant. Hence $\log \left(p_{1} / p_{2}\right)$ is constant on the line through $D$ and $E$. In a similar way any perpendicular on $A C$ defines a constant $\log \left(p_{1} / p_{3}\right)$. Where the two perpendiculars cross, we find $(u, v)$. With proper choice of the anchor points, the relationship between the logratios and the coordinates of the point $(u, v)$ is very simple: $u=\log \left(p_{2} / p_{1}\right)$ and $v=\left(2 \log \left(p_{3} / p_{1}\right)-\right.$ $\left.\log \left(p_{2} / p_{1}\right)\right) / \sqrt{3}$. See the Appendix for details.

We do not use log-ratios themselves as new coordinates, but linear combinations of them. In fact they are scaled version of the default isometric log-ratio transform (ilr). The default choice of ilr in the R package compositions (van den Boogaart et al., 2014) gives $u^{\prime}=\log \left(p_{2} / p_{1}\right) / \sqrt{2}$ and $v^{\prime}=\left(2 \log \left(p_{3} / p_{1}\right)-\log \left(p_{2} / p_{1}\right)\right) / \sqrt{6}$, so $u=u^{\prime} \sqrt{2}$ and $v=v^{\prime} \sqrt{2}$.

We use oblique axes for the log-ratios to map from a composition, constrained to the simplex, to an unbounded two-dimensional space. In fact we effectively stretch the sides of the familiar ternary triangle without limits.

In Figure 3 we show four ternary diagrams for different selections of metals in the Kola data set in the R package StatDA (Filzmoser, 2013). The corresponding TrioScale plots appear in Figure 4.

Note that the anchor points are not shown, because they play no role anymore in the alternative interpretation of TrioScale. Only axes for the three log-ratios have to be drawn. We can move them to a parallel position any distance we like. In Figure 4 the axes are identical for all three


Figure 2: Diagram supporting the TrioScale mapping. The difference $|A E|^{2}-|B E|^{2}$ is constant for any position of $E$ on the perpendicular through $D$. If $\log p_{2}=\exp \left(-|B E|^{2}\right)$ and $\log p_{1}=\exp \left(-|A E|^{2}\right)$, the $\log -r a t i o \log \left(p_{2} / p_{1}\right)$ is constant on the perpendicular.
log-ratios, in all four panels. To see more detail, it is convenient to limit the range of each axis and position it close to the cloud of points. That has been done in Figure 5.


Figure 3: Kola data set (moss). Four selected triples of metals. Presentation as standard ternary plots.


Figure 4: Kola data set (moss). Four selected triples of metals. Presentation in TrioScale coordinates. The three axes represent logs of ratios of the metals indicated at both of their ends.


Figure 5: Kola data set (moss). Four selected triples of metals. Presentation in TrioScale coordinates. The three axes represent logs of ratios of the metals indicated at both of their ends. Contours of bivariate normal distributions added.

## 3 Parametric distributions

After TrioScale mapping, observations form a cloud of points in a two-dimensional Euclidean space. Familiar multivariate statistical techniques can be used to describe their distribution. The bivariate normal distribution is an obvious choice. Figure 5 shows that this might be a reasonable idea, even for data that show rather extreme patterns in the ternary diagram.

We believe that statistical analysis should be done in the TrioScale space. Nevertheless, it is illustrative to map the elliptic contours back to the ternary diagram. This is done in Figure 6. Notice that we take account of the Jacobian of the transformation. That increases the density near the corners and edges of the ternary diagram.

In the literature on distributions for compositional data one can find proposals for normal distribution with and without scaling by the Jacobian (Egozcue et al., 2003). The ellipse contours obtained from the R package compositions (van den Boogaart et al., 2014) does not account for the Jacobian. See the next section for the technical details of transforming a density numerically.

Apparently the bivariate normal distribution does not give a perfect fit. In most of the triples of metals there is a strong suggestion of a skewed shape. This becomes more noticeable if we compute the singular value decomposition and plot the distributions of the scores on both axes, as is done in Figure 7. In three of four cases we see a skew distribution on the first singular vector. We will not discuss possibilities for complicated parametric models, as we feel that this would take us too far afield. Instead we present non-parametric density estimation in the section that follows.

## 4 The scatterplot smoother

If a scatterplot contains many points, it is hard to judge their density well. Eilers and Goeman (2004) proposed a fast solution. They form a two-dimensional histogram and smooth it along the rows and then along the columns (or the other way around, with identical results). The procedure is very fast en even when written in pure R it can handle hundreds of thousand of data points in less than a second. Application in TrioScale coordinates is straightforward. Figure 8 shows results


Figure 6: Transformed elliptic contours of the bivariate normal distribution (in TrioScale space), shown in the ternary diagram.


Figure 7: Histograms of the elements of first (SV 1) and second (SV 2) singular vectors.


Figure 8: Kola data set (moss). Four selected triples of metals. Presentation in TrioScale coordinates. The three axes represent logs of ratios of the metals indicated at both of their ends. Scatter plot smoothing.
for the chosen subsets of the Kola moss data.
The scatterplot smoother gives a pleasant display and a better view of the distribution of the data. In Figure 8 the skewness of three of the data clouds is clearly visible Also there is a suggestion, in the two right panels, of a mixture of two convex-shaped components.

To transform a smooth density back to a ternary diagram we use the following procedure, which is inspired by the way van den Boogaart and Tolosana-Delgado (2013) compute contours for the Dirichlet distribution.

- Save the smooth density (in TrioScale coordinates) in the matrix $H$, together with the corresponding $u$ and $v$ grids.
- Form a fine grid in the $(x, y)$ ternary coordinates.
- Transform to compositions.
- Select only those grid points where each composition is larger than 0.001 and smaller than 0.999 .
- For each grid point do the following
- Transform the compositions to Trioscale coordinates.
- Determine the corresponding element of $H$. Let its value be $f$.
- Compute the determinant of the Jacobian matrix of the transformation from $(x, y)$ to $(u, v)$. Let that be $g$.
- Save the product of $f$ and $g$ in a matrix, index by $x$ and $y$.

The result is a matrix with proper density values for a ternary display. It can be shown as an image and enhanced with contours, if desired, see Figure 9.


Figure 9: Kola data set (moss). Four selected triples of metals. Back-transformed non-parametric densities in ternary diagrams.

## 5 Discussion

We have shown how to use TrioScale to change from bounded compositions to unbounded coordinates in a plane and do density smoothing there. Numerical back-transformation gives proper density estimates in the ternary diagram. The unpleasant artifacts of direct use of kernel estimators, which do not respect boundaries, are avoided.

Visualization is not only math but also psychology. This is even more important in the ternary diagram. Close to corners and edges densities can become quite narrow. Whereas contour lines help interpretation in TrioScale coordinates, where they generally are reasonably spaced, they can get too close together in the ternary diagram and become a nuisance.

Our color scale is linear in the value of the density. In some case it might be preferably to use a square root (or other power smaller then 1) to enhance low-density regions. We use rev(heat.colors(100)) for the scale. It runs from white to red, which is pleasant when printed on paper. R itself and the package colorspace offer many other color scales.

The annotation of the axes may need more attention. A slight problem is that round numbers for the natural logarithms are not easy to interpret on the probability scale, as they correspond to $e \approx 2.72, e^{2} \approx 7.39$ et cetera, which are not round numbers. Logarithms to base 10 lead to more familiar, but rather large ratios. A compromise might be to use logarithms to base 2. Numbers like 2,4 and 8 are easy to handle in one's head.

## References

Carr D., ported by Nicholas Lewin-Koh, Martin Maechler and contains copies of lattice function written by Deepayan Sarkar (2014). hexbin: Hexagonal Binning Routines. $R$ package version 1.27.0. http://CRAN.R-project.org/package=hexbin
de Rooij, M. (2009). Ideal point discriminant analysis with a special emphasis on visualisation. Psychometrika 74, pp. 317-330.
de Rooij, M. and Eilers, P.H.C. (2013) TrioScale: A new diagram for compositional data. Proceedings of the 5th International Workshop on Compositional Data Analysis, pp. 18-27.

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

Eilers, P.H.C and Goeman, J. (2004). Enhancing Scatterplots with Smoothed Densities. Bioinformatics 20, pp. 623-628.
Filzmoser, P. (2013). StatDA: Statistical Analysis for Environmental Data. R package version 1.6.7. http://CRAN.R-project.org/package=StatDA

Hamilton N. (2015). ggtern: An Extension to 'ggplot2', for the Creation of Ternary Diagrams. $R$ package version 1.0.5.0. http://CRAN.R-project.org/package=ggtern
van den Boogaart, K.G. and Tolosana-Delgado, R. (2013) Analyzing Compositional Data with R. Springer.
van den Boogaart, K.G., Tolosana R. and Bren, M. (2014). compositions: Compositional Data Analysis. $R$ package version 1.40-1. http://CRAN.R-project.org/package=compositions

## Appendix

How do we get from a composition vector $p$ to its position in the TrioScale plane? Let this point have coordinates $(u, v)$ and let the coordinates of anchor point $j$ be $\left(u_{j}, v_{j}\right)$. Then we have that

$$
\begin{equation*}
\log p_{j}=-\left(u-u_{j}\right)^{2}-\left(v-v_{j}\right)^{2}+\log c, \tag{1}
\end{equation*}
$$

where $c$ is the normalizing constant. Thus

$$
\begin{equation*}
\log \frac{p_{j}}{p_{k}}=\left(u-u_{k}\right)^{2}+\left(v-v_{k}\right)^{2}-\left(u-u_{j}\right)^{2}-\left(v-v_{j}\right)^{2} . \tag{2}
\end{equation*}
$$

After simplification we get one linear equation in $u$ and $v$ :

$$
\begin{equation*}
2\left(u_{j}-u_{k}\right) u+2\left(v_{j}-v_{k}\right) v=r_{j k}=\log \frac{p_{k}}{p_{j}}-u_{k}^{2}+u_{j}^{2}-v_{k}^{2}+v_{j}^{2} . \tag{3}
\end{equation*}
$$

From the log-ratio of $p_{2}$ and $p_{1}$, and that of $p_{3}$ and $p_{1}$ follow two linear equations. The solution for $(u, v)$ is found by solving

$$
\left[\begin{array}{ll}
u_{2}-u_{1} & v_{2}-v_{1}  \tag{4}\\
u_{3}-u_{1} & v_{3}-v_{1}
\end{array}\right]\left[\begin{array}{l}
u \\
v
\end{array}\right]=\left[\begin{array}{l}
r_{21} / 2 \\
r_{31} / 2
\end{array}\right] .
$$

We can simplify these equations by a proper choice of the coordinates of the anchor points. Take $u_{1}=v_{1}=v_{2}=0, u_{2}=2, u_{3}=1$ and $v_{3}=\sqrt{3}$. Also we can switch to $u^{\prime}=2 u$ and $v^{\prime}=2 v$ Then we get

$$
\left[\begin{array}{cc}
2 & 0  \tag{5}\\
1 & \sqrt{3}
\end{array}\right]\left[\begin{array}{l}
u^{\prime} \\
v^{\prime}
\end{array}\right]=\left[\begin{array}{l}
r_{21} \\
r_{31}
\end{array}\right]=\left[\begin{array}{l}
\log \left(p_{2} / p_{1}\right)-4 \\
\log \left(p_{3} / p_{1}\right)-4
\end{array}\right] .
$$

We can drop the -4 terms: they lead to a shift in the $(u, v)$ plane and absolute coordinates are not relevant for a graphical display. After inverting the matrix on the left-hand side we find the simple transformations:

$$
\begin{equation*}
u=\log \left(p_{2} / p_{1}\right) ; \quad v=2 \log \left(p_{3} / p_{1}\right) / \sqrt{3}-\log \left(p_{2} / p_{1}\right) / \sqrt{3}, \tag{6}
\end{equation*}
$$

where we have dropped the accents.
These simple linear equations in the log-ratios of proportions look similar to those for the coordinates in the ternary diagram

$$
\begin{equation*}
x=p_{2}+p_{3} / 2 ; \quad y=p_{3} \sqrt{3} / 2 \tag{7}
\end{equation*}
$$

which are expressed in the proportions themselves.
Notice that the $\log$-ratios $\log \left(p_{2} / p_{1}\right)$ and $\log \left(p_{3} / p_{1}\right)$ define lines in the plane and the crossing of the lines determines a position in that plane. This is different from using the log-ratios themselves as new coordinates. Conversely, from the point in the $u, v$ plane follow log-ratios, but its coordinates are not log-ratios themselves.

When we transform from $(x, y)$ to $(u, v)$ and $f(u, v)$ is the density on the $(u, v)$ plane, we have that density $\breve{f}(x, y)=f(u, v) g(x, y)$. Here $g=|G|$, where the bars indicate the determinant and $G$ is the matrix of partial derivatives:

$$
G=\left(\begin{array}{ll}
\frac{\partial u}{\partial x} & \frac{\partial u}{\partial y}  \tag{8}\\
\frac{\partial v}{\partial x} & \frac{\partial v}{\partial y}
\end{array}\right)=\left(\begin{array}{cc}
\frac{1}{p_{1}}+\frac{1}{p_{2}} & \frac{1}{\sqrt{3}}\left(\frac{1}{p_{1}}-\frac{1}{p_{2}}\right) \\
\frac{1}{\sqrt{3}}\left(\frac{1}{p_{1}}-\frac{1}{p_{2}}\right) & \frac{1}{3}\left(\frac{1}{p_{1}}+\frac{1}{p_{2}}+\frac{4}{p_{3}}\right)
\end{array}\right) .
$$

It may seem a bit unusual to express the partial derivatives in terms of the proportions instead of $x$ and $y$, but the expressions are simple and we compute $p$ anyway. After some algebraic manipulation we find that

$$
\begin{equation*}
|G|=\frac{4}{3}\left(\frac{1}{p_{1} p_{2}}+\frac{1}{p_{1} p_{3}}+\frac{1}{p_{2} p_{3}}\right) \tag{9}
\end{equation*}
$$

# A compositional approach to allele sharing analysis 

I. Galván ${ }^{1}$, J. Graffelman ${ }^{2}$ and C. Barcelol ${ }^{1}$<br>${ }^{1}$ Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona; ivan.galvan@udg.edu; carles.barcelo@udg.edu<br>${ }^{2}$ Department of Statistics and Operations Research, Universitat Politècnica de Catalunya; jan.graffelman@upc.edu


#### Abstract

Relatedness is of great interest in population-based genetic association studies. These studies search for genetic factors related to disease. Many statistical methods used in population-based genetic association studies (such as standard regression models, t-tests, logistic regression) assume that the observations (individuals) are independent. These techniques can fail if independence is not satisfied.

Allele-sharing is a powerful data analysis technique for analyzing the degree of dependence between individuals. Two individuals can share 0,1 or 2 alleles for any genetic marker. This sharing may be assessed for alleles identical by state (IBS) or identical by descent (IBD). Starting from IBS alleles, it is possible to detect the type of relationship of a pair of individuals by using graphical methods. Typical allele-sharing analysis consists of plotting the fraction of loci sharing 2 IBS alleles versus the fraction of sharing 0 IBS alleles (Rosenberg, 2006). Compositional data analysis can be applied to allele-sharing analysis because the proportions of sharing 0,1 or 2 IBS alleles (denoted by $p_{0}, p_{1}$ and $p_{2}$ ) can be considered as a 3 -part-composition. This paper provides a graphical method to detect family relationships by plotting the isometric log-ratio transformation of $p_{0}, p_{1}$ and $p_{2}$ in ilr-coordinates. On the other hand, the probabilities of sharing 0,1 or 2 IBD alleles (denoted by $k_{0}, k_{1}, k_{2}$ ), which are termed Cotterman's coefficients (Cotterman, 1941), depend on the relatedness: monozygotic twins, full-siblings, parent-offspring, avuncular, first cousins, etc. It is possible to infer the type of family relationship of a pair of individuals by using maximum likelihood methods (Thompson, 1975; 1991). As a result, the estimated vector of $\hat{\boldsymbol{k}}=\left(\hat{k}_{0}, \hat{k}_{1}, \hat{k}_{2}\right)$ for each pair of individuals can be considered as a 3-part-composition and plotted in a ternary diagram to identify the degree of relatedness.

An R package has been developed for the study of genetic relatedness based on genetic markers such as microsatellites and single nucleotide polymorphisms from human populations, and is used for the computations and graphics of this contribution.


Key words: Allele sharing, identical by state, identical by descent, Cotterman's coefficients, ternary diagram, isometric log-ratio transformation.

## 1 Introduction

The application of statistics in genetics and molecular biology has become an active field of research over the last decades. Studies of family relationships in genetic data analysis are crucial in population-based genetic association studies (Foulkes, 2009). The main aim of research in these association studies is to find genetic factors related to disease. These studies assume that individuals from human populations are independent. The dependence between individuals, e.g. the presence of related individuals in a database, can invalidate the statistical methods applied in association studies such as regression models or t-tests. Thus, it is important to detect the degree of relatedness of a pair of individuals in the database. This can help to avoid such dependence by removing one individual of the detected pair.

Genetic datasets are composed of genetic markers that are helpful to find the possible DNA regions related to a disease of interest such as cancer. Single nucleotide polymorphisms (SNPs) and microsatellites or short tandem repeats (STRs) are common genetic markers in population-based genetic association studies (Laird and Lange, 2011). SNPs are common throughout the human genome, they occur once every 300 nucleotides on average in the DNA sequences formed by the base pairs adenine (A), cytosine (C), guanine (G) and thymine (T). It is well known that there are approximately 10 million SNPs in the human genome. Consequently, SNPs are mostly used for large scale genetic association studies. We give an example of a SNP for three individuals who have the following DNA sequences on a pair of chromosomes at a specific locus: $\mathrm{ID} 1=(\mathrm{CCGATC}$, CCAATC), ID2=(CCGATC, CCGATC) and ID3=(CCAATC, CCAATC). Note that for the first individual the sequences differ only at the third base pair for the alleles G and A and for the second and the third individual the same alleles appear at the third base. Thus, the third position is a SNP, in this case a G/A polymorphism. The three individuals have a SNP coded by the genotypes GA, GG and AA, respectively, at this specific locus.

On the other hand, microsatellites are short DNA sequences which are repeated. The length of the repeated DNA sequences is constant for each STR and ranges from 2 to 6 nucleotides. Unrelated individuals have genetic variability because their alleles of determined regions of DNA vary. As a result, microsatellites are very powerful to distinguish each individual from the population due to the presence of genetic variability between individuals. Particularly, the number of the repeated sequences across STRs varies between unrelated individuals. For this reason, they are also used for forensic DNA studies. There are two ways to code an STR: by recording the total size in base pairs of the repeating sequences; or by considering only the number of repeats of a particular sequence. For instance, an individual has the following DNA sequences on a pair of chromosomes at a specific locus: $\operatorname{ID1}=($ ATTATTATTATT, ATTATTATTCCC). This is a trinucleotide repeat ATT that can be coded as an STR of $(4,3)$ repeats or as an STR of size $(12,9)$.

Allele sharing analysis is a classical technique for analyzing the degree of dependence between individuals. Two individuals can share 0,1 or 2 alleles for any genetic marker. The larger the number of shared alleles between a pair of individuals across genetic markers, the more likely they are to be closely related. Thus, individuals from the same family share on average more alleles than unrelated individuals. Allele sharing can be considered by state if the DNA composition of the alleles is identical but the alleles do not come from a common ancestor (identical by state, IBS). It can also be considered by descent if the alleles originate from a common ancestor (identical by descent, IBD). We consider both methods for relatedness research in the rest of this paper.

This paper is organized as follows. Section 2 gives an overview of the IBS allele sharing analysis and the application of ilr-coordinates. Section 3 presents the basic principles of the IBD allele sharing and the representation of the family relationships in a ternary diagram. Sections 2 and 3 treat examples of IBD and IBS studies with microsatellite and SNP data. An R package that can simulate data for the studies is discussed (still in development). Finally, Section 4 summarizes the principal conclusions of this article.

## 2 Identical by state studies

IBS studies ignore if the alleles of any pair of individuals are derived from a common ancestor. The IBS sharing of a pair of individuals can be calculated from the genotype data. Then, two individuals share 0 IBS alleles if they have no alleles in common (e.g. AA and GT); share 1 IBS allele if one individual has only a single allele in common with the other individual (e.g. AA and AT or AA and TA; the position of the alleles is irrelevant), and 2 IBS alleles if they have identical genotypes (e.g. AA and AA). Occasionally, the number of shared alleles may be missing (NA) if some individual has missing genotyping information (e.g. AA and NA or NA and NA).

This approach is usually considered for all the pairs of individuals from a human population across genetic markers. Then, for each pair of individuals we have a vector of 0,1 or 2 shared alleles as large as the number of genetic markers in the database. Consequently, it is possible to build a vector $\boldsymbol{p}$ of the proportions of shared alleles $(0,1,2)$ for each pair of individuals denoted $\boldsymbol{p}=\left(p_{0}, p_{1}, p_{2}\right)$ respectively. Classical IBS allele sharing consists of plotting the proportion of sharing 2 IBS alleles $\left(p_{2}\right)$ versus the proportion of sharing 0 IBS alleles $\left(p_{0}\right)$ for all pairs of individuals from a given human population (Rosenberg, 2006). This graphical method is powerful to detect family relationships by observing the pairs of individuals with higher values of $p_{2}$. It is known that family relationships of degree zero, that is, monozygotic twins (MZ), usually have values of $p_{2}$ close to 1 . Parentoffspring pairs (PO) usually have values of $p_{0}$ close to 0 . Full-siblings (FS), half-siblings (HS), avuncular (AV), grandparent-grandchild (GG) are also family relationships that can be detected by this graphical method. Unrelated individuals (UN) usually have higher values of $p_{0}$. However, the plot $p_{2}$ versus $p_{0}$ ignores the constraint $p_{0}+p_{1}+p_{2}=1$ and the relative information of the component $p_{1}$.

For this reason, we propose the isometric log-ratio (ilr) transformation (Egozcue et al., 2003) of the vector $\left(p_{0}, p_{1}, p_{2}\right)$ in order to preserve the relative information of the 3 parts. The resultant ilr-coordinates can be plotted to detect family relationships in IBS studies.

By construction, the first ilr-coordinate $\left(z_{1}\right)$ interprets the balance between $p_{2}$ and $p_{0}$. Note that this coordinate captures the information of the graphical method explained before by Rosenberg. The second ilr-coordinate $\left(z_{2}\right)$ corresponds to the balance between $p_{0}, p_{2}$ and $p_{1}$. The ilr-coordinates are defined as follows:

$$
\left\{\begin{array}{l}
z_{1}=\frac{1}{\sqrt{2}} \ln \left(\frac{p_{2}}{p_{0}}\right)  \tag{1}\\
z_{2}=\frac{\sqrt{2}}{\sqrt{3}} \ln \left(\frac{\sqrt{p_{0} p_{2}}}{p_{1}}\right)
\end{array}\right.
$$

MZ, PO, FS, HS, AV, GG pairs have higher values of $z_{1}$ whereas unrelated individuals have lower values of $z_{1}$.

### 2.1 Example

We present an R package called IBS.IBD.studies. This package contains a sample from the Maya population of 25 individuals extracted from a world-wide database from the Noah A. Rosenberg Research lab at Stanford University (Rosenberg, 2002; 2006). This world-wide database is derived from the Human Genome Diversity Cell Line Panel (HGDP) (Cavalli-Sforza, 2005). For each individual, the sample from the Maya population includes 5 columns with their individual code number assigned by the HGDP (ID), the population code number assigned by Rosenberg's lab (Pop.Code), the population name (Pop.Nam), the geographic information (Geographic) and the region of the population (Region). The genetic information consists of 377 microsatellites (STRs) labeled by their respective "locus names" (D12S1638, D14S1007, ...). Table 1 shows a glance at the database of the Maya population.

Table 1: A glance at the Maya population.

| ID | Pop.Code | Pop.Nam | Geographic | Region | D12S1638 | D14S1007 | D9S1779 | D9S1825 | $\ldots$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 854 | 86 | Maya | Mexico | AMERICA | 120 | 130 | 142 | 129 | $\ldots$ |
| 854 | 86 | Maya | Mexico | AMERICA | 120 | 124 | 124 | 129 | $\ldots$ |
| 855 | 86 | Maya | Mexico | AMERICA | 126 | NA | 124 | 137 | $\ldots$ |
| 855 | 86 | Maya | Mexico | AMERICA | 120 | NA | 124 | 129 | $\ldots$ |
| 856 | 86 | Maya | Mexico | AMERICA | 126 | 130 | 124 | 135 | $\ldots$ |
| 856 | 86 | Maya | Mexico | AMERICA | 126 | 128 | 124 | 129 | $\ldots$ |
| 857 | 86 | Maya | Mexico | AMERICA | 126 | 128 | 124 | 135 | $\ldots$ |
| 857 | 86 | Maya | Mexico | AMERICA | 126 | 128 | 124 | 129 | $\ldots$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |  |

Each individual from the Maya population is listed in two consecutive lines. For instance, the genotype for the individual $\mathrm{ID}=854$ for the STR D12S1638 is $(120,120)$. The allele " 120 " indicates the total size in base pairs of the repeating DNA sequence. An individual whose genotyping information is missing, is coded by NA (not available).

We use the functions allelesharing() and percentages () from the IBS.IBD.studies R package. The first function computes the shared IBS alleles for each genetic marker and for each pair of individuals. Using percentages(), the proportions of sharing 0,1 or 2 IBS alleles $\left(p_{0}, p_{1}, p_{2}\right)$ for each pair of individuals from the sample are obtained.

Figure 1 plots the fraction of loci sharing $2 \operatorname{IBS}$ alleles $\left(p_{2}\right)$ versus the fraction of sharing 0 IBS alleles $\left(p_{0}\right)$ for all pairs of individuals from the Maya population. The family relationships documented by Rosenberg (2006) are represented in different colours. Observe that outlying individuals correspond to family relationships of the first degree (PO and FS) and the second degree (AV or GG); relationships of the third degree such as first cousins (FC) are more difficult to detect and are mixed with unrelated individuals (UN).

Figure 1: Plot of the fraction of loci sharing 2 IBS alleles $\left(p_{2}\right)$ versus the fraction of sharing 0 IBS alleles ( $p_{0}$ ) for all pairs of individuals from the Maya population.


Figure 2 shows the representation in ilr-coordinates of all pairs of individuals of the Maya population. The ilr-coordinates are built by using the function ilr() from the R package compositions (Boogaart et al., 2013). According to the first ilr-coordinate $\left(z_{1}\right)$, the distance between the two PO pairs and the UN pairs from the Maya population equals three units, whereas the distance between the FS pair and the UN pairs is approximately one unit. The second ilr-coordinate also discriminates the related individuals from the unrelated individuals. According to the second ilrcoordinate $\left(z_{2}\right)$, the distance between PO and UN is two units; the distance between FS and UN is one unit. The AV, GG, FC pairs are the points which are closest to the unrelated pairs in the data.

Figure 2: Representation of all pairs of individuals from the Maya population in ilr-coordinates.


Comparing Figures 1 and 2, we note that, PO pairs are more outlying in ilr-coordinates.

## 3 Identical by descent studies

We have shown that IBS studies offer graphical methods to detect relatedness between individuals from human populations. Here, we present IBD studies in order to identify accurately which type of family relationship belongs to each pair of individuals. The degree of relatedness can be inferred by considering the number of IBD alleles shared. The probabilities of sharing 0,1 or 2 IBD alleles are called Cotterman's coefficients (Cotterman, 1941) and are denoted by $k_{0}, k_{1}$ and $k_{2}$ respectively. These probabilities depend on the relatedness: monozygotic twins (MZ), parent-offspring (PO), full-sibs (FS), half-sibs (HS), avuncular (AV), grandparent-grandchild (GG), first cousins (FC) or unrelated individuals (UN) and are presented in Figure 3 (left). Note that HS, AV and GG have exactly the same Cotterman coefficients and is not possible to distinguish them, unless we build a pedigree tree for the given human population under study. The set of IBD probabilities has the 2-simplex as its domain. For this reason, it is possible to represent all the family relationships in a ternary diagram as is shown in Figure 3 (right).

Figure 3: Left: Cotterman's coefficients for the different type of family relationship and degree of relatedness. Right: Representation of the Cotterman coefficients in a ternary diagram.

| Type of Relative | Degree | $k_{0}$ | $k_{1}$ | $k_{2}$ |
| :--- | :---: | :---: | :---: | :---: |
| Monozygothic twins | 0 | 0 | 0 | 1 |
| Parent-offspring | 1 | 0 | 1 | 0 |
| Full-siblings | 1 | $1 / 4$ | $1 / 2$ | $1 / 4$ |
| Half-siblings/ avuncular/ <br> grandchild-grandparent | 2 | $1 / 2$ | $1 / 2$ | 0 |
| First cousins |  |  |  |  |
| Unrelated | $\infty$ | $3 / 4$ | $1 / 4$ | 0 |



In practice, genetic data contains information for estimating the IBD probabilities. However, these probabilities depend on the genotypes and the allele frequencies from the human population under study. For this reason, a good estimation procedure of the Cotterman coefficients is needed. If the estimated IBD probabilities are close or coincide with the theoretical Cotterman coefficients from Table 3 (left) for a given relationship, then that relationship is inferred. Many articles have been found in the literature on the estimation of Cotterman's coefficients. Maximum likelihood estimation is the best estimation method (Milligan, 2003; Weir, 2006). We do not detail the steps of this method in this contribution, nevertheless, a detailed explanation of the Cotterman coefficients and their estimation can be found in CoDaWork 2015 by Graffelman and Galván (2015). In this article, we compute the maximum likelihood estimates of $k_{0}, k_{1}$ and $k_{2}$ by using the function cotterman() from the IBS.IBD.studies R package. This function uses the optimization routines from the Rsolnp R package (Ghalanos and Theussl, 2014).

Once the estimates of $\left(k_{0}, k_{1}, k_{2}\right)$ are obtained, graphical methods such as the plot of $k_{1}$ versus $k_{0}$ (Nembot-Simo et al., 2013) or $k_{2}$ versus $k_{1}$ (Moltke and Albrechtsen, 2014) are commonly used to identify the degree of relatedness. These plots separate the related individuals from the unrelated individuals, however, they ignore the relative information of the remaining part of the 3 -part-composition $\boldsymbol{k}=\left(k_{0}, k_{1}, k_{2}\right)$. Therefore, it seems logical to plot the Cotterman coefficients in a ternary diagram as an additional graphical method in order to identify relatedness. This way, the information of the IBD probabilities is preserved for all pairs of individuals and the family relationships are inferred by comparing the estimates with the theoretical values of $k_{0}, k_{1}$ and $k_{2}$ plotted in Figure 3 (right). The Cotterman coefficients can be plotted in a ternary diagram by using the function ternaryDiag () from the robCompositions R package (Templ et al., 2014) as shown by the examples below.

### 3.1 Examples

In his Section we use simulated and empirical datasets and plot them in a ternary diagram in order to identify relationships. The functions $\operatorname{simSNP}()$ and children() from the R package IBS.IBD.studies can be used to simulate genetic marker data with given family relationships. First, we generate a sample of 20 unrelated individuals with 1000 genetic markers. The function simSNP() simulates random single nucleotide polymorphisms (SNPs), giving categorical variables with the three genotypes $\mathrm{AA}, \mathrm{AB}$ and BB . All SNPs have minor allele frequency of 0.5 . SNPs are simulated independently and under the assumption of Hardy-Weinberg Equilibrium: $p_{A}^{2}+2 p_{A} p_{B}+$ $p_{B}^{2}=1$ (Foulkes, 2009). Thus, each SNP is a random sample of a multinomial distribution of size 20 (the number of unrelated individuals). The theoretical genotype probabilities of $\mathrm{AA}, \mathrm{AB}$ and BB are $0.25,0.5$ and 0.25 respectively. Each individual is labeled by 'ID' as shown in Table 2.

Table 2: A glance at the simulated dataset.

| Individuals | SNP1 | SNP2 | SNP3 | SNP4 | $\cdots$ | SNP1000 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| ID1 | AA | AB | AB | AB | $\cdots$ | BB |
| ID2 | AA | AB | AB | BB | $\cdots$ | AB |
| ID3 | BB | AB | AB | AA | $\cdots$ | BB |
| ID4 | AB | BB | AB | AA | $\cdots$ | AB |
| ID5 | AB | AB | BB | AA | $\cdots$ | BB |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| ID20 | BB | AA | AB | AA | $\cdots$ | BB |

Once the sample is generated, we use the function children() to build a pedigree tree as follows. Because we know that a child always has received one allele from the father and one allele from the mother, the function children() chooses one allele randomly across SNPs from the individual 'ID1' and one allele from the individual 'ID2' to generate a child. This new individual is labeled by 'ID21A'. Analogously, we produce another child labeled by 'ID21B'. Thus, 'ID21A' and 'ID21B' are a full-siblings pair. We complete the pedigree tree by generating a child (labeled by 'ID22') from the individuals 'ID2' and 'ID3' in order to originate two half-siblings pairs, which correspond to the pairs 'ID21A'-'ID22' and 'ID21B'-'ID22'. An additional family relationship was created, a duplicated individual of 'ID15' (labeled by 'ID23'); hence, 'ID15' and 'ID23' represent a monozygotic twin pair. Thus, the new simulated population consists of 24 individuals. The simulated relationships are composed of one MZ pair, six PO pairs, one FS pair and two HS pairs as shown in Figure 4.

Figure 4: The simulated family relationships. Left: a pedigree tree consisting of six PO pairs, one FS pair and two HS pairs. Right: a MZ pair.


We again study the Maya population (Section 2.1) as an empirical example.
Figure 5 shows the ternary diagrams of the estimated Cotterman coefficients ( $k_{0}, k_{1}, k_{2}$ ) for all pairs of individuals of the simulated human population and the Maya population respectively. All the family relationships of degree zero, one and two are close to the theoretical probabilities described in Figure 3. However, relationships of degree three such as FC in Figure 5 (right) are difficult to discriminate from UN or AV.

Figure 5: Ternary diagrams of the estimated Cotterman coefficients $\left(k_{0}, k_{1}, k_{2}\right)$ for all pairs of individuals from the simulated data (left) and the Maya population (right).


## 4 Conclusion

This article is focused on the annotation of the relatedness between individuals in genetic data. We stress the importance of the detection of family relationships as a tool for quality control in population-based genetic association studies. The statistical methods used in these studies can fail if the dependence between individuals is not documented. We have shown two classical approaches in relatedness research and we have applied tools from compositional data for the identification of family relationships.

First, the IBS allele sharing analysis provides a graphical tool for detecting related individuals. This graph plots the proportion of sharing 2 IBS alleles versus the proportion of sharing 0 IBS alleles. We propose an additional graphical method by using the isometric log-ratio transformation of the vector of proportions of sharing 0,1 or 2 IBS alleles. We plot the ilr-coordinates for all pairs of individuals and use this plot to identify relationships.

Finally, the IBD allele sharing analysis offers an accurate estimation of relatedness by using Cotterman's coefficients. Plots of $k_{1}$ versus $k_{0}$ or $k_{2}$ versus $k_{1}$ are used to represent graphically the family relationships from a population. These plots ignore the constraint $k_{0}+k_{1}+k_{2}=1$ and we state that ternary diagrams may be useful to identify family relationships. The theoretical values of $k_{0}, k_{1}$ and $k_{2}$ form reference points in the ternary diagram for the standard relationships.

## Acknowledgements

This study was supported by grant CODARSS MTM2012-33236 of the Spanish Ministry of Education and Science.

## References

Boogart K.G.v.d., Tolosona R. and Bren M. (2013). Compositions: Compositional Data Analysis. R package version 1.40.0.

Cavalli-Sforza L.L. (2005). The Human Genome Diversity Project: past, present and future. Nature Reviews, Genetics 6 pp. 333-340.

Cotterman C.W. (1941). Relative and human genetic analysis. Sci Monthly 53, pp. 227-234.
Egozcue J.J. et al. (2003). Isometric Logratio Transformations for Compositional Data Analysis. Mathematical Geology 35 (3) pp. 279-300.

Foulkes, A.S. (2009). Applied Statistical Genetics with R. Springer.
Ghalanos, A. and Theussl, S. (2014). Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method. R package version 1.15.

Graffelman J. and Galván I. (2015). An application of the isometric log-ratio transformation in relatedness research. The 6th International Workshop on Compositional Data Analysis, CoDaWork 2015.

Laird N.M. and Lange C. (2011). The fundamentals of modern statistical genetics. Springer.
Milligan B.G.(2003). Maximum-likelihood estimation of relatedness. Genetics, 163 pp. 1153-67.
Moltke I. and Albrechtsen A (2014). RelateAdmix: a software tool for estimating relatedness between admixed individuals. Bioinformatics, $30 \mathrm{pp} .1027-8$.

Nembot-Simo A., Graham J. and McNeney B. (2013). CrypticIBDcheck: an R package for checking cryptic relatedness in nominally unrelated individuals. Source Code for Biology and Medicine, 8:5.

Rosenberg N.A. (2002) Rosenberg lab at Stanford University: http://www.stanford.edu/group/rosenberglab/diversity.html

Rosenberg N.A. (2006). Standardized Subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, Accounting for Atypical and Duplicated Samples and Pairs of Close Relatives. Annals of Human Genetics 70, pp. 841-847.

Templ A., Hron K. and Filzmoser P. (2014). robCompositions: Robust Estimation for Compositional Data. R package version 1.8.0.

Thompson E.A. (1975). Estimation of pairwise relationships. Annals of Human Genetics, 39 pp. 173-188.

Thompson E.A. (1991). Estimation of relationships from genetic data. In: Rao CR, Chakraborty R (Eds.) Handbook of Statistics. Elsevier Science: Amsterdam, Vol 8, pp. 255-269.

Weir B.S., Anderson A.D., Hepler A.B. (2006). Genetic relatedness analysis: modern data and new challenges. Nature Reviews, Genetics 7 pp. 771-780.

# Compositional uncertainty in high-throughput sequencing data analysis should not be ignored. 

Gregory B. Gloor ${ }^{1}$, Andrew D. Fernandes ${ }^{2}$, Jean M. Macklaim ${ }^{1}$, and Michael Vu ${ }^{1}$<br>${ }^{1}$ The University of Western Ontario, London, Canada; ggloor@uwo.ca<br>${ }^{2}$ YouKaryote Genomics, London, ON, Canada


#### Abstract

High throughput sequencing is used for many different types of experimental analyses including 16 S rRNA gene sequencing, transcriptomics and metagenomics. All of these experiments represent random sampling of sequence fragments derived from an underlying unknown distribution of fragments and attempt to determine which, if any, of the sampled fragments are differentially abundant between predefined groups. These data typically are composed of thousands of fragments in only a small number of samples. The number of sequence reads binned to each fragment represents a single estimate of the underlying frequency of the fragment in each sample. The values obtained are compositional because the total number of sequences obtained is trivially determined by the instrument itself and provides no information about the number of random fragments in the input pool. In addition, in many instances, the investigator is interested in the relative, and not the absolute abundance, of the fragments.

We show that the variation due to sampling alone (technical variation) in compositional datasets derived from high-throughput sequencing is large and inversely related to the number of reads mapping to a fragment. Ignoring this technical variation can lead to false positive inference regardless if the data is treated as a composition or not. We find that a two-step procedure that incorporates a Baysian estimate of fragment abundance along with the centred log-ratio transformation markedly improves specificity with no loss of sensitivity.


Key words: Bayesian estimation, centred log-ratio, transcriptomics, metagenomics, 16 S rRNA gene sequencing, ALDEx2

## 1 Introduction

High throughput sequencing studies, that generate as outputs thousands to billions of sequence tags, are becoming the norm in the life sciences. That these experiments generate compositional data can be understood with two statements. First, the total number of sequence tags obtained in an experiment are of no importance. Second, the sequence tags are binned into features where the difference between features is exponential and best explained by ratios. These features can represent genes as in 16 S rRNA gene sequencing, transcriptomics and metagenomics or singlenucleotide variant abundances after differential growth experiments. The experimentalist in these experiments is interested in knowing which features, if any, are differentially abundant between two or more distinct groups. Furthermore, all experiments of this type explicitly or implicitly examine sub-compositions. Finally, each individual experimental design is analyzed using different sets of underlying assumptions that are derived from historical dogma, despite having the same underlying data structure. Fernandes (2014) demonstrated that tools developed for one experimental design (e.g. RNA-Seq) do not translate well to other experimental designs (e.g. 16S RNA gene sequencing).

These data are necessarily sparse, and complex. There are often hundreds or thousands of features, and the expense of these experiments prevents the collection of sufficient sequence tags to ensure that all features are covered by at least one sequence tag. Thus, the treatment of features with zero counts is a pervasive problem when treating these data as compositions (Lovell et al. 2011). It is assumed that features with zero counts across all samples are removed because they are uninformative. For the remainder where one approach is to delete features where one or more samples have zero counts (Lovell et al. 2011; Lovell et al. 2014). This removes the problem of zero count features at the expense of potentially excluding the most important features from consideration. Another approach, is to replace the zero counts with an expected value calculated in some way. Several approaches with differing underlying assumptions are in use, and Martín-Fernández(2014) suggested that a Bayesian-Laplace approach to be the most reasonable. Regardless of the method used to treat zero count features, these analyses use maximum-likelihood approaches to determine feature abundance prior to analyses.

We have found that the variation due to sampling alone (technical variation) in compositional datasets derived from high-throughput sequencing is large and inversely related to the number of reads mapping to a fragment (Fernandes et al. 2013). Ignoring this technical variation can lead to false positive inferences regarding differential abundance if the data are not treated as compositions. We have found that a two-step procedure incorporating a Bayesian estimate of feature abundance along with analyses conducted after centred-log-ratio transformation markedly improves specificity with no loss of sensitivity, and that the increase in specificity derived almost entirely from the exclusion of low-count (including zero count) features (Fernandes et al. 2014).

Our paper explores how the analyses differ when the value of zero is assigned using different approaches with, and without Bayesian estimation of the technical variation. Our initial work showed that a uniform prior added to all values was able to encompass the estimated technical variation in a sparse dataset (Fernandes et al. 2013). However, we observed that this approach slightly overestimated technical variation of low count and zero count features, suggesting that this approach had less than optimum power.

We will compare uniform priors that replace 0 , uniform priors added to all values, and the prior estimation methods from the zCompositions package that produce non-uniform estimates of the actual zero value. We will examine a real differential growth experimental dataset for which an objective standard of truth is known. We argue that these results are generalizable across other datasets including RNA-seq datasets and 16 S rRNA gene sequencing experiments.

## 2 Results

High throughput sequencing is a technology that delivers thousands to millions of reads that correspond uniquely to genes or other features in a genome, or to bins that represent sequence variants. Figure 1A shows several different study designs that are common in the literature. Regardless of design a very large number of molecules, shown in the orange box in Figure 1A are randomly sampled to produce a library that is then sequenced. The sequencing instrument delivers a much smaller random sample of the actual input. In addition, the act of sequencing converts the data from unconstrained to constrained proportional data because the instrument can deliver a fixed upper number of sequence reads. This hard upper bound means that all such analyses generate compositional data regardless of the actual study design. These experiments all aim to ask the question, "what gene or feature has a different abundance between groups A and B?"


Figure 1: High-throughput sequencing affects the shape of the data. Panel A illustrates the workflow by which high throughput sequencing samples the DNA or RNA from an environment. There are many more molecules that are sampled than can be incorporated into the library, or that can be sequenced on the instrument. The capacity of the instrument itself determines the number of reads observed. The orange box shows the representative number of molecules in the initial sample, and the blue box shows the maximum number of reads that are obtained from the instrument. These reads are assigned to features such as genes or operational taxonomic units or other bins, and a table of the reads per feature is output. Panel B illustrates how the data is distorted during the process. The input DNA or RNA usually has no fixed sum and is randomly sampled sequentially during the library preparation and sequencing steps. The output from the instrument is proportional because the instrument can deliver only a fixed upper limit of reads, regardless of the number of molecules in the input. Panel B. 1 shows the number of reads in the input tube for 15 steps where the open square and circular features are held at a constant number and the black feature is increasing in abundance by 2-fold each step. Panel B. 2 shows the output in proportions (or ppm) after random sampling to a constant sum, as occurs on the sequencer. Panel B. 3 shows the shape of the data following centre log-ratio transformation.

Figure 1B shows how sequencing distorts the data. Many processes examined by high throughput sequencing can be thought of as linear compositional processes. Consider a mixture of many distinctive molecules in vector $x=\left[x_{1}, x_{2}, \ldots x_{n}\right]$ over time or space increments $i$. For each increment we can determine the abundance of each molecule using Equation 1:

$$
\begin{equation*}
x_{i}=x_{0} \times 2^{(\lambda i)} \tag{1}
\end{equation*}
$$

where $\lambda$ is the growth rate. If $\lambda=0$ for all but one of the members of vector $x$ and $\lambda=1$ for one member, then one member will double in abundance at each increment and all remaining members will be unchanged at each increment. Figure 1B. 1 shows such a thought experiment where the values are plotted as counts of molecules. Producing a library and sequencing it generates a set of counts per gene that are scaled by the maximum number of reads delivered by the machine. In other words, the counts for gene $x_{i}$ are per-gene probabilities $p_{i}$ and are formally equivalent to a random multivariate Poisson sample of the original group of DNA molecules. We model this process by sampling from the Dirichlet distribution according to Equation 2:

$$
\begin{equation*}
\left[p_{1}, p_{2}, \ldots p_{n}\right] \sim \operatorname{Dirichlet}\left[x_{1}, x_{2}, \ldots x_{n}\right] \tag{2}
\end{equation*}
$$

A single Dirichlet instance done in this way generates a single Bayesian estimate of the underlying posterior probabilities for each feature, and multiple samples generate a full posterior distribution (Fernandes et al. 2013). Panel 1B. 2 shows the posterior values for a single Dirichlet instance from the counts in Panel 1B.1. Here we can see that the constant sum constraint that results from the finite read limit of the instrument severely distorts the underlying shape of the data. Figure 1B. 3 demonstrates that applying the standard centred log-ratio transform of Aitchison (1986) to the vector of probabilities $p$ in Panel 1B. 1

$$
\begin{equation*}
\operatorname{clr}(p)=\left[\log _{2} \frac{p_{1}}{g(p)}, \log _{2} \frac{p_{2}}{g(p)}, \ldots \log _{2} \frac{p_{n}}{g(p)}\right] \tag{3}
\end{equation*}
$$

reconstitutes the essential shape of the data, with the actual data points for now some variability because of random sampling. In this equation, $g(p)$ is the geometric mean of the vector $p$. This transformation is convenient because it reconstitutes the essential shape of the original data, and because there is a one to one mapping between the values in the original and in the transformed dataset. Furthermore, this transformation is easily interpretable because it is simply a ratio between the abundance of a gene or feature in the sample and the average abundance of all genes or features in the sample.

### 2.1 Data from high-throughput sequencing is highly variable

Data from high throughput sequencing experiments are often thought of as point estimates despite being random samples of the input molecules, and despite several experiments showing that sequencing the same DNA library will produce somewhat different count tables at the same sequencing depth (Marioni et al. 2008; Bottomly et al. 2011). Figure 2 shows an example of this variability. Marioni (2008) did an experiment where two aliquots of the same RNA-seq library were run in duplicate, and the resulting reads were mapped to the $>20000$ genes in the human genome. Replicate runs did not return exactly the same number of reads per gene: for example, when the genes in one replicate contained zero counts, the same genes in the other replicate often had non-0 reads. This imprecision extended across the range of per-gene counts as shown for a few replicate read values in Figure 2. This imprecision is proportionally larger for small count values, and smaller for large count values. For example, the range of counts observed in replicate B when genes in replicate A contain one count span the range of 0-14 in this example: a difference of over 10 -fold. By comparison, when genes in replicate A contain 64 counts the corresponding genes in replicate B span counts from 38-91: a difference of less than $50 \%$. See Figure 1 of Fernandes (2013)


Figure 2: Technical replicate variation of an RNA-seq experiment showing the the count observed for features from a one replicate of an RNA-seq dataset as a function of the count for the same feature in the other replicate. The base value count is give above each histogram, and the location of this count in the histogram is shown as the black bar.
for a demonstration that the proportional error does indeed span the entire range of expression values in this dataset.

The imprecision can be modelled by sampling instances from a Dirichlet distibution (Fernandes et al. 2013; Fernandes et al. 2014) as in Equation 2. Figure 3 shows quantile-quantile plots comparing the distribution of true technical variation to the distribution of technical variation estimated by drawing instances from the Dirichlet distribution. That is for a vector of counts $x, x_{\text {Dir }}=$ Dirichlet $[x] \times \sum x$. Sampling multiple Dirichlet instances returns a distribution of the posterior probabilities of each feature in the vector $x$, and conserves probability. These plots show that the Dir instances slightly over-estimate the mean abundance when the counts are low but approach the observed distribution when counts are in double digits. We conclude that drawing Dirichlet instances is an acceptable proxy for true technical replication in these datasets.

### 2.2 False positive results because of unaccounted variation

One problem when analyzing such data is that the available datasets - whether derived from 16 S rRNA gene sequencing, transcriptomics, or other experiments - are exploratory and so generally lack a standard of truth. This makes it difficult to develop and test tools without modelling a dataset. While modelled datasets have some allure because the parameters can be closely controlled, we prefer to examine the behaviour of different approaches in real biological datasets because they often have unanticipated error and less predictable behaviour than modelled datasets.

McMurrough (2014) generated a selective growth (selex) dataset for which a standard of truth can be inferred. This dataset compares the growth of a set of 1600 sequence variants in the I-LtrI endonuclease under two conditions. The first condition is a non-restrictive condition where the growth of all variants is unconstrained. The second condition is restrictive for growth, unless the I-LtrI endonuclease is active and can cleave and inactivate the gene encoding $C c d b$, a DNA gyrase toxin. The gyrase toxin is dose-dependent so cleavage of a fraction of the plasmids containing the gene confers slower growth (Smith and Maxwell 2006), and under the conditions of the assay, the toxin would be bacteriostatic if no cleavage occurred. Thus in this experimental design the


Figure 3: QQ plots comparing the actual distribution of technical variation to the distributions of 100 Dirichlet instances for genes in the Marioni dataset that have the counts indicated above each plot. The red line illustrates equivalence.
difference between inactive variants between the two conditions would be one of dilution alone, and no variant should become less abundant during the experiment. Variants that cleave the toxin gene would confer a growth advantage, and would become more abundant over the time of the assay. Furthermore, McMurrough (2014) showed that the in vitro enzymatic activity of the endonuclease is strongly correlated with the output of the selective growth experiment.

The abundance of each variant in the mixture can be modelled by Equation 1. At time zero if each variant is contained in vector $n_{0}=\left[n_{1}, n_{2}, n_{3} \ldots n_{1600}\right]$, over time increments, the change in abundance in the non-selected growth conditions can be modelled with $\lambda=1$ and the variation in $\lambda$ being small. The experimental conditions allowed for approximately 16 doublings, or time increments. Therefore at the last increment of the non-selected time series, we anticipate that the initial relationships between the abundances of each of the 1600 variants will be essentially unchanged. In contrast, the selected variants are under strongly differential selection. Here the most active variants will have $\lambda=1$, that is, these variants grow at the same rate in the selected and unselected conditions. The least active variants will have $\lambda=0$, that is, these variants will not change in actual abundance during selection, but will become relatively less abundant when compared to their active counterparts. Inactive variants are known to be by far the most abundant in each sample. Intermediate positive values of $\lambda$ are expected, and no negative values are expected. Finally, it is possible for individual samples to demonstrate differences in apparent $\lambda$ under selection. This can occur if a variant is partially active, and cleaves all the toxin genes in a particular cell by chance. This event is heritable and so would allow the cell to grow at the maximum rate. Thus, the sample in which this occurred would have an apparent increase in $\lambda$ for that variant in that sample.

The question we wish to address with this dataset is: can we identify from the growth experiment alone which variants are likely to be active? Active variants will have had a maximum of 16 cell doublings becoming much more abundant, inactive variants will stay at the same abundance and variants with partial activity will become only somewhat more abundant. In addition, we wanted to know the effect on our inference of the different approaches to estimating the zero values. We first examined the dataset using a biplot to show the relationship between the samples and the variants.


Figure 4: Biplot showing the relationship between samples and variables in the selex dataset. Zero values were adjusted using the count zero multiplicative approach using the zCompositions $R$ package, and the biplot was generated using the compositions R package. The vast majority of the variables cluster around the centre of the dataset. The red arrows to the left show that the 7 non-selected replicates are very similar, and the selected replicates on the right exhibit some variability. The differences between samples are driven by variation in a small number of variables. In this dataset, component 1 explains $52.4 \%$ of the variability and component 2 explains $10.4 \%$.

A compositional biplot generated with the compositions R package (van den Boogaart and TolosanaDelgado 2008) following zero replacement using the CZM approach from the zCompositions R package (Palarea-Albaladejo and Martín-Fernández 2015) is shown in Figure 4. The first two components of this biplot explained $52.4 \%$ and $10.4 \%$ of the variance in the data. The selected and non-selected samples separate clearly on the first component, and this separation is driven largely by the abundance of the variants on the right side which McMurrough (2014) demonstrated to be the highly active variants. However, it is difficult to quantitate the magnitude of their abundance change from this analysis. This figure shows that the non-selected samples, which cluster on the left side, are essentially redundant since the links between them are exceedingly short. The selected samples are much more diverse, and some of the links from the origin are nearly orthogonal. Inspection of the data finds that this diversity is driven by only a few variables, and that these variables separate the X1 and X2 sample sets: these sets are from identical experiments performed with different batches of the same cell type. Examinination of the underlying count table shows that these variants are indeed strongly different in abundance between the X1 and X2 sets. For example the I:E:V:E variable has 17933 reads in sample X2_DS but has zero reads in sample X1_DS. This is an example of a single stochastic event that conferred a growth advantage to this variant in
this sample. It is important to note that this large stochastic variation has important consequences when examining datasets because such variation is not unusual in real biological datasets.

### 2.2.1 Not accounting for sampling results in many false positive identifications

The selex dataset is unique because we have a validated truth for some of the features that differentiate the conditions (McMurrough et al. 2014). In this dataset we have unambiguously biochemically identified variants that are active and those that are not. Based on this prior information, we expect that approximately 60 variants would exhibit substantial activity in this assay, and so substantial deviation from this number would indicate many false positive results.

One approach that is widely used in the literature is to reduce the values from the count table to proportions or normalized counts through Maximum Likelihood approaches, then to conduct univariate statistical tests for each variant, and (sometimes) to correct for multiple testing (see for example the 16 S rRNA analysis methods in Hsiao (2013)). These approaches often treat zero values as actual zeros, thus making no adjustments. Applying this simple method using an unpaired Wilcoxon test, and applying the Benjamini-Hochberg correction (Benjamini and Hochberg 1995) to the resulting P values reveals that 1593 of 1600 variants are identified as having a differential abundance between the selected and non-selected conditions with an adjusted P value cutoff of 0.05 . This is clearly at odds with the known biology of the underlying dataset.

A potentially more rigorous, yet still simple approach is to adjust the zero values in this dataset using one of the available methods that are implemented in the $z$ Compositions R package (PalareaAlbaladejo and Martín-Fernández 2015) and then to treat the data as compositions by applying Equation 3 before performing univariate statistical tests. Recall from Figure 1 that this transformation recapitulates the essential shape of the data, and there is a one to one mapping of variant counts to centred log-ratio values. The range of values for zero replacement by different methods are given in the Prior column of Table 1. Interestingly, three of these zero correction methods returned values greater than 1 for some of the zero values, this likely was a result of the very large difference between the selected and non-selected count values in the two groups. In addition, we applied two other approaches to deal with zero values. The first, labeled uniform replacement, replaces all zero values with 0.5 but does not adjust other values in the dataset. This is akin to adding a pseudo count to zero values. The second labelled uniform prior, applies a uniform prior adjustment to all values in the selex dataset. For this we use the minimally informative Jeffrey's prior of 0.5 . The Comp 1 and Comp 2 columns in Table 1 show the percentage variation explained by a clr compositional bipolar using each of these zero adjustments in the first two principle components. Only the biplot that used the square-root Bayesian multiplicative method appears to result in a transformation that explains substantially less of the variation in the dataset.

Table 1: Numbers of distinguishing features identified in the selex growth experiment observed with different approaches to assign prior expectations to zero count features.

| Prior assignment | Prior | Comp 1 | Comp 2 | Point | Dir |
| :--- | :---: | :---: | :---: | :---: | ---: |
| Count zero | $0.325452-0.325910$ | 0.524 | 0.104 | 874 | 91 |
| Geometric Bayesian | $0.061279-4.890273$ | 0.504 | 0.108 | 355 | 82 |
| Square root | $0.006854-3.102299$ | 0.452 | 0.118 | 1008 | DNR |
| Bayes-Laplace | $0.030497-4.883747$ | 0.480 | 0.108 | 435 | 133 |
| Uniform replacement | 0.5 | 0.556 | 0.098 | 958 | 74 |
| Uniform Prior | 0.5 | 0.528 | 0.102 | 868 | 84 |

This approach was tested by adjusting zero values in the count table by the identified method, closing the resulting vectors, applying the centred log-ratio transform to each sample and then subjecting the features to unpaired Wilcoxon tests. Again P values were adjusted using the BenjaminiHochberg method and an adjusted P value of 0.05 was used as the threshold for significance. Table

1-Point shows the results of this approach. Here we see that all of the methods substantially improve upon the naive approach, with between one-quarter and two-thirds of the variables being identified as differential. In this dataset, the square root Bayesian multiplicative method provides the largest number of positive identifications, and the Geometric Bayesian multiplicative correction provides the smallest number of positive identifications, although no method is able to strongly distinguish the known small number of true positives from a much larger number of false positives.

### 2.2.2 Accounting for sampling reduces false positive identifications

One substantial shortcoming of these approaches is that the inherent technical variation in the dataset that results from random sampling during sample preparation and sequencing as shown in Figure 2 is not taken into account. It is becoming an accepted practice to account for the sampling using Dirichlet multinomial mixture models, where each sample is represented by a vector of probabilities, rather than point estimates (Holmes et al. 2012). For example, Ding (2014) recently used this approach to partition microbiomes into different community states in a robust manner. We applied this approach by generating 128 Dir instances of the selex dataset using Equation 2 with a uniform prior of 0.5 , and then conducted per-variant Wilcoxon tests on each instance. The mean Benjamini-Hochberg adjusted P value for each variant was tabulated, and again the cutoff used was an adjusted P value of 0.5 . Surprisingly, this approach, which takes into account the inferred technical variation, again resulted in 1593 of the 1600 variants as being differentially abundant between the selected and non-selected groups. This is more than the 868 variants detected when variation was not taken into account but the centred log-ratio transform was applied, and equivalent to the naive method accounting for neither variation nor the compositional nature of the data. Thus, simple averaging across inferred the technical variates is not sufficient to screen out false positive variants in this dataset.

Finally, we used an approach that combined generating 128 Dirichlet instances of the data followed by applying the centred log-ratio transformation before conducting significance tests. This method is implemented in the ALDEx2 R package for the analysis of high throughput sequencing datasets (Fernandes et al. 2013; Fernandes et al. 2014), and is available at Bioconductor.

As implemented, the ALDEx2 package uses the uniform zero replacement value of 0.5 . One purpose of this investigation was to determine if using one of the more rigorous zero replacement models from the zCompositions package would increase our selectivity because, as shown in Table 1, these adjustments output non-uniform estimates of the underlying value of zero based on abundances of the same feature in different samples.

We applied the same seven methods to adjust the value of zero in the selex dataset, and an overview of the results are shown in the Dir column of Table 1. We again used Wilcoxon tests on the two groups and corrected the resulting P values using the Benjamini-Hochberg approach. Significance was assumed if the mean adjusted P value across all 128 instances was less than 0.05 . In this analysis the substituted values of zero in the adjusted datasets serve as prior estimates of the range of values that zero could assume in each of the Dirichlet instances. The square root Bayesian multiplicative approach was incompatible with generating Dirichlet instances because many of the prior values that replaced zero generated Dirichlet posterior estimates that were not distinguishable from zero. Modelling uniform priors indicated that this occurred when the prior for zero was less than approximately 0.05 . The remaining six approaches were compatible with the approach, and resulted in substantially smaller numbers of variants being identified as significantly different between the selected and non-selected groups. In this analysis, the Geometric Bayesian multiplicative, uniform replacement and uniform prior approaches were approximately similar, the count zero multiplicative approaches was nearly as selective, and the Bayes-Laplace approach was least selective.

Figure 5 shows a variance-variance plot of the output from an analysis using the uniform prior replacement with a value of 0.5 . Note that in this plot the vast majority of variants have an


Figure 5: Variance-variance plot showing the median maximum centred log-ratio scaled difference within each group plotted vs. median between group difference for each variant. Dotted lines represent the approximate location of effect sizes, which is calculated as the median between to within group difference. Variants are coloured if their activity was validated in vitro, have a star if they failed to grow reliably in individual culture in vitro. Variants that exhibit a significant increase in abundance using the Wilcoxon test with a mean Benjamini-Hochberg adjusted P value of $>0.05$ are shown as large grey dots. The analysis was done with a uniform prior of 0.5 applied to the dataset. Also shown are the six variants that were outliers on component 2 of the clr biplot in Figure 4
estimated between group difference of approximately zero, that only a small number have a positive between group difference, and no variants have a strong negative between group difference. This fits with the experimental design where variants could increase in abundance if the endonuclease was active, but not decrease in abundance if it was not. In this plot the variants with a mean BenjaminiHochberg adjusted P value determined by an unpaired Wilcoxon test are indicated by the large grey dots. Variants that were tested for enzymatic activity in vitro are indicated by coloured central dots. Variants that had near wild type enzymatic activity in vitro are in the sector marked as $>8$. There were four variants that had partial enzymatic activity in vitro. Many variants were tested for growth in pure culture. Variants AEAE, SEGE, ADGD and GDAD exhibited variable, partial growth under these conditions, with the GDAD variant exhibiting the weakest growth. Thus, there is a strong relationship between the observed results in this experiment, and the results observed in vitro.

There is remarkable concordance between the data viewed in this way, and the same data viewed as a point estimate in the compositional biplot. The biplot shows that the most distinguishing variants between the selected and non-selected groups, i.e., the variants that drive the separation on principle component 1, are those in the upper left quadrant of Figure 5. In addition, the variants that drive the separation on principle component 2, are those that exhibit the largest
within-condition difference. For example, the GEME, AEME, PDME, IEVE, PEQD, and DEAD variants that were strongly separated on component 2 on the biplot, are among those with the largest within group difference on the variance-variance plot.


Figure 6: Variance-variance plots showing how the prior values for zero determined by the count zero multiplicative (CZM), geometric Bayesian multiplicative (GBM) and Bayes-Laplace (BL) methods alter the variation of the data. Red dots represent those that are called differential, black dots are not differential and the lines represent effect sizes of 1 and -1 . The cutoff used was a Benjamini-Hochberg adjusted P value of 0.05 from an unpaired Wilcoxon test.

Finally, we examined the effect of the different zero replacement methods on the shape of the variance-variance plot to determine why these different approaches deliver slightly different results after Dirichlet sampling log-ratio transformation. As shown in Figure 6 all the prior estimation methods delivered similar differences between conditions for the true positive variants. These all exhibited an increase in abundance of about $2^{16}$ relative to their mean abundance in the unselected group. In particular, the CZM plot was remarkably similar to the plot that used a uniform prior of 0.5 , with the major difference between the two approaches being a slight broadening of the withingroup difference. This is perhaps not surprising since the prior values of zero using this approach are non-uniform in a narrow range of near 0.325. In contrast the non-uniform prior values for zero count variants from both the GBM and BL ranged over much larger values. The vast majority of values were between zero and one, but the GBM method had an average of 197.4 zero replacements that were greater than one, and the BL had an average of 55 replacements that were greater than one. Examination of the variance-variance plots of these two approaches showed that betweengroup difference for many variants was not significant, but tended to be strongly negative. This result is incompatible with the known biology of the experiment, where no variant is expected to become less abundant than average in the selected dataset. Therefore, in this dataset, the geometric Bayesian multiplicative and the Bayes-Laplace substitution methods are distorting the underlying data. This distortion likely contributes to the greater number of variants identified as significantly different between the selected and non-selected groups.

## 3 Discussion

High throughput sequencing datasets are very different from other types of datasets to which compositional approaches are often applied. High throughput sequencing datasets result from random sampling of a large number of DNA fragments, and the act of sequencing these DNA fragments on the instrument results in data that has the constant sum constraint. The estimation of the true abundance of genes or features with low counts exhibits a very large proportional error. However the posterior distribution of the underlying abundance of each feature can be estimated by generating multiple instances of the data by sampling from a Dirichlet distribution. These datasets are necessarily very sparse, but in many cases the sparsity is informative. For example, in 16 S rRNA gene sequencing it is difficult to argue that a particular taxonomic group would never
be observed if we generated sufficient sequencing reads. As another example, gene expression is stochastic, and the number of transcripts for a given gene is observed not to be zero when large populations of cells are sampled, even if the gene in question is 'not expressed' (Munsky et al. 2012).

The centred log-ratio approach is intrinsically attractive in a biological context for two reasons. First, it can be intuitively explained to biologists as being similar to quantitative PCR, a familiar technique where the ratio between the gene of interest and a gene assumed to be at a constant level is determined. The centred log-ratio approach merely extends this analogy to the ratio between the gene of interest and all other genes in the system. Second, biologists understand that many of the processes that they study, cell growth, enzyme kinetics, etc, are exponential processes. Less well understood is that the underlying data is not 'set in stone' but actually represents a snapshot of what would have been observed had the experiment been done again.

A common criticism of using log-ratio approaches when analyzing such sparse data is the problem of zero observed counts. Structural zeros, those features that contain zero in every sample, are always excluded, and do not cause problems. However, count zeros that occur in one condition but not the other are problematic because log-ratio transformations cannot be performed when the underlying data contains one or more features with a zero value (Aitchison 1986). Much work has been put into this problem because of the prevalence of features with values of zero are common in many kinds of datasets. Several approaches have been developed to determine the best point estimate of the actual underlying value of zero in these datasets (Pawlowsky-Glahn et al. 2015), and they are implemented in zCompositons R package (Palarea-Albaladejo and Martín-Fernández 2015). Less work has been done modelling this in a Bayesian framework where the distribution of probable values for each variable are taken into account.

Here we have examined the effect of using various approaches to estimating the value of zero on both point estimates and Bayesian distributions derived from Dirichlet multinomial sampling. We have found that point estimates, whether modelled as proportions or centre log-ratio transformed values, cannot distinguish features that differ between conditions. We found that estimating the technical variation alone is also unsuitable. However, the combination of estimating technical variation and the centre log-ratio transformation provides a large increase in selectivity. We further observe that methods that generate priors in a narrow range give outputs that closely mimic a dataset derived from a differential growth experiment, and that methods that generate priors with broad ranges generate posterior distributions that are different from the known underlying distribution.

The selex dataset is an extreme example of the type of data that is analyzed by high throughput sequencing. It has a small number of features that exhibit a marked difference in abundance between conditions, and is very sparse. Other experimental designs will have much smaller difference in abundance of features. For example, in the case of RNA-seq it is more common to examine differential abundance of a small number of genes that are themselves relatively rare in the cell. This would be akin to comparing steps 1 and 2 in Figure 1B.1. In this simple case, any approach would likely give reasonable answers. However, comparing gene expression between cells from different tissues, or gene expression in RNA from environmental samples, would introduce more extreme distortions in the underlying data and could give false positive and false negative results (Fernandes et al. 2013; Macklaim et al. 2013; Fernandes et al. 2014). In the case of 16S rRNA gene sequencing experiments, it is likely that many conditions would have wildly divergent underlying abundances because bacterial growth is an exponential process, and such samples are more difficult to analyze.

## Acknowledgements

This work was supported by a grant to Greg Gloor by the National Science and Engineering Research Council of Canada.

## References

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Chapman \& Hall.
Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 289-300.
Bottomly, D., N. A. R. Walter, J. E. Hunter, P. Darakjian, S. Kawane, K. J. Buck, R. P. Searles, M. Mooney, S. K. McWeeney, and R. Hitzemann (2011). Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-seq and microarrays. PLoS One 6(3), e17820.

Ding, T. and P. D. Schloss (2014, May). Dynamics and associations of microbial community types across the human body. Nature 509(7500), 357-60.
Fernandes, A. D., J. M. Macklaim, T. Linn, G. Reid, and G. B. Gloor (2013, July). ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-seq. PLoS ONE 8(7), e67019.
Fernandes, A. D., J. N. Reid, J. M. Macklaim, T. A. McMurrough, D. R. Edgell, and G. B. Gloor (2014). Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16s rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome 2, 15.
Holmes, I., K. Harris, and C. Quince (2012). Dirichlet multinomial mixtures: generative models for microbial metagenomics. PLoS One 7(2), e30126.
Hsiao, E. Y., S. W. McBride, S. Hsien, G. Sharon, E. R. Hyde, T. McCue, J. A. Codelli, J. Chow, S. E. Reisman, J. F. Petrosino, P. H. Patterson, and S. K. Mazmanian (2013, Dec). Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. Cell 155(7), 1451-63.
Lovell, D., W. Müller, J. Taylor, A. Zwart, C. Helliwell, V. Pawlowsky-Glahn, and A. Buccianti (2011). Proportions, percentages, ppm: do the molecular biosciences treat compositional data right? Compositional Data Analysis: Theory and Applications, 193-207.
Lovell, D., V. Pawlowsky-Glahn, J. J. Egozcue, S. Marguerat, and J. Bähler (2014). Proportionality: a valid alternative to correlation for relative data. bioRxiv, 008417.
Macklaim, M. J., D. A. Fernandes, M. J. Di Bella, J.-A. Hammond, G. Reid, and G. B. Gloor (2013). Comparative meta-RNA-seq of the vaginal microbiota and differential expression by Lactobacillus iners in health and dysbiosis. Microbiome 1, 15.
Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad (2008, Sep). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18(9), 1509-17.
Martín-Fernández, J.-A., K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo (2014). Bayesian-multiplicative treatment of count zeros in compositional data sets. Statistical Modelling doi:10.1177/1471082X14535524, 1:25.
McMurrough, T. A., R. J. Dickson, S. M. F. Thibert, G. B. Gloor, and D. R. Edgell (2014, Jun). Control of catalytic efficiency by a coevolving network of catalytic and noncatalytic residues. Proc Natl Acad Sci U S A 111(23), E2376-83.

Munsky, B., G. Neuert, and A. van Oudenaarden (2012, Apr). Using gene expression noise to understand gene regulation. Science 336(6078), 183-7.
Palarea-Albaladejo, J. and J. A. Martín-Fernández (2015). zcompositions - r package for multivariate imputation of left-censored data under a compositional approach. Chemometrics and Intelligent Laboratory Systems 143(0), 85-96.
Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2015). Modeling and Analysis of Compositional Data. John Wiley \& Sons.

Smith, A. B. and A. Maxwell (2006). A strand-passage conformation of dna gyrase is required to allow the bacterial toxin, ccdb, to access its binding site. Nucleic Acids Res 34 (17), 4667-76.
van den Boogaart, K. G. and R. Tolosana-Delgado (2008). "compositions": A unified R package to analyze compositional data. Computers \& Geosciences $34(4), 320-338$.

# Microbial Denitrification and the Hardy-Weinberg law 

J. Graffelman ${ }^{1}$ and R. De Wit ${ }^{2}$<br>${ }^{1}$ Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain; jan.graffelman@upc.edu<br>${ }^{2}$ Centre for Marine Biodiversity, Exploitation and Conservation (MARBEC), CNRS \& Université de Montpellier, Montpellier, France; rutger.de-wit@umontpellier.fr


#### Abstract

An application of compositional data analysis in the fields of bio-geochemistry and microbial ecology is presented. In the surface layer of aquatic sediments, bacterial processes contribute to the conversion of bound nitrogen into nitrogen gas. These natural processes represent a way to reduce the negative impacts of nutrient over-enrichment (eutrophication) and, therefore, have been the subject of many biogeochemical and ecological studies. By using isotope tracking experiments with ${ }^{15} \mathrm{~N}$ the contribution of the different bacterial processes to the nitrogen conversion can be investigated. Massspectrometry measurements of nitrogen gas produced by incubated sediments samples produces compositional data sets consisting of 3-part compositions of nitrogen gas containing ${ }^{28} N_{2},{ }^{29} N_{2}$ and ${ }^{30} N_{2}$. Under the null hypothesis of $100 \%$ classical microbial denitrification, these 3 parts are expected to occur in Hardy-Weinberg proportions. We use tools from compositional data analysis such as ternary diagrams and isometric log-ratio transformations of the data to test this hypothesis. We found evidence that classical denitrification, although probably predominant, coexisted with other processes like Anammox coupled to ammonium oxidizers, and had to reject the Hardy-Weinberg law as a model for the data.


Key words: Eutrophication, isotope tracking, Anammox process, nitrate reduction, ammonium oxidation, isometric log-ratio transformation, Hardy-Weinberg equilibrium.

## 1 Introduction

This paper concerns an application of compositional data analysis in the fields of bio-geochemistry and microbial ecology. In the surface layer of aquatic sediments, bacterial processes contribute to the conversion of bound nitrogen into nitrogen gas. These natural processes represent a way to reduce the negative impacts of nutrient over-enrichment (eutrophication) and, therefore, have been the subject of many biogeochemical and ecological studies (Deutsch and others, 2010; Kana and others, 1998; Nielsen, 1992). These processes are typically studied experimentally by using isotope tracking. It consists of adding nitrate to the water column that is labeled with the heavy isotope ${ }^{15} \mathrm{~N}$ (i.e., ${ }^{15} \mathrm{NO}_{3}^{-}$). The naturally predominant $N$ isotope is ${ }^{14} \mathrm{~N}$ at a proportion of $(1-p)=0.9963$, and the natural abundance of ${ }^{15} N$ is $p=0.0037$. The experimental incubation of such a system results in the production of three different forms of dinitrogen gas $\left({ }^{28} N_{2},{ }^{29} N_{2}\right.$ and $\left.{ }^{30} N_{2}\right)$ that differ in mass according their isotopic composition $\left({ }^{14} N\right.$ and ${ }^{15} N$ ) and the produced amounts can be measured by using mass-spectrometry. The reaction zone is anoxic and located just below the oxic-anoxic transition zone in the sediment. The ${ }^{15} \mathrm{NO}_{3}^{-}$arrives in this zone by molecular diffusion from the water column, while the ${ }^{14} \mathrm{NO}_{3}^{-}$arrives by diffusion from the surficial oxic sediment layers, where it is produced by nitrification, and partly also from the water column (when the ${ }^{15} \mathrm{~N}$ labeling of this nitrate is not complete). Hence by knowing the proportion of ${ }^{15} \mathrm{~N}$ labeled nitrate in the water column it becomes possible to quantify the link between nitrification and denitrification. The relative amounts of the three types of $N_{2}$ form a composition that can be represented in a ternary diagram (Aitchison, 2003). This diagram can be interpreted in the same manner as is done in genetics (Graffelman \& Morales-Camarena, 2008), and gives rise to a molecular formulation of the Hardy-Weinberg principle. The Hardy-Weinberg law is a cornerstone principle of modern genetics, and states that under random combination of the alleles of a diallelic marker, the genotype frequencies $A A, A B$ and $B B$ will occur in the relative proportions $p^{2}, 2 p(1-p)$ and $(1-p)^{2}$, where $p$ is the allele frequency of the A allele. The genotype frequencies can be represented in a ternary diagram with the three genotypes as vertices. The base of the triangle is $0-1$ axis for the allele frequency, and a parabola inside the diagram represents the Hardy-Weinberg proportions. In the context of our microbiological experiments, The Hardy-Weinberg parabola is predicted for the classical denitrification of nitrate, because the random combination of ${ }^{14} \mathrm{NO}_{3}^{-}$and ${ }^{15} \mathrm{NO}_{3}^{-}$implies that ${ }^{28} N_{2},{ }^{29} N_{2}$ and ${ }^{30} N_{2}$ will be produced in the relative proportions $(1-p)^{2}, 2 p(1-p)$ and $p^{2}$ where $p$ is the relative proportion of ${ }^{15} \mathrm{NO}_{3}^{-}$in the reaction zone for the denitrification. A ternary diagram can be constructed with the three types of nitrogen gas as vertices. The base of the triangle is a 0-1 axis for the relative amount of ${ }^{15} \mathrm{~N}$, and a parabola inside the diagram represents the expected Hardy-Weinberg proportions under classical denitrification, as depicted in Figure 1.

The Hardy-Weinberg proportions for the isotopic composition of dinitrogen are, however, disturbed by the anaerobic oxidation of ammonium into nitrogen with nitrite $\left(\mathrm{NO}_{2}^{-}\right)$as the electron donor, the so-called Anammox process (Kuenen, 2008). The nitrite for this process may be provided either by i) the reduction of nitrate into nitrite in the anoxic zone (Anammox coupled to nitrate reduction) or by ii) the oxidation of ammonium into nitrite (Anammox coupled to ammonium oxidizers). When classical denitrification coexists with i) the isotope compositions will fall above the parabola, and when it coexists with ii) they are below the parabola. Thus, the following processes play a role in the denitrification process:

- Classical denitrification

$$
{ }^{14} \mathrm{NO}_{3}^{-}+{ }^{15} \mathrm{NO}_{3}^{-} \longrightarrow{ }^{28} \mathrm{~N}_{2},{ }^{29} \mathrm{~N}_{2},{ }^{30} \mathrm{~N}_{2} .
$$

The random combination of ${ }^{14} \mathrm{NO}_{3}^{-}$and ${ }^{15} \mathrm{NO}_{3}^{-}$implies that ${ }^{28} \mathrm{~N}_{2},{ }^{29} \mathrm{~N}_{2}$ and ${ }^{30} \mathrm{~N}_{2}$ will be produced in the relative proportions $(1-p)^{2}, 2 p(1-p)$ and $p^{2}$ where $p$ is the relative proportion of ${ }^{15} \mathrm{NO}_{3}^{-}$in the reaction zone for the denitrification.

- The Anammox process

$$
{ }^{14,15} \mathrm{NO}_{2}^{-}+{ }^{14} \mathrm{NH}_{4}^{+} \longrightarrow 2 \mathrm{H}_{2} \mathrm{O}+{ }^{28} \mathrm{~N}_{2},{ }^{29} \mathrm{~N}_{2} .
$$

In this process, the random combination of ${ }^{14} \mathrm{NO}_{2}^{-}$and ${ }^{15} \mathrm{NO}_{2}^{-}$implies that ${ }^{28} \mathrm{~N}_{2}$ and ${ }^{29} \mathrm{~N}_{2}$ will be produced in the relative proportions $(1-p)$ and $p$ respectively. The nitrite used in the Anammox process can come from two sources. It can be produced by ammonium oxidation $\left({ }^{14} \mathrm{NH}_{4}^{+} \longrightarrow{ }^{14} \mathrm{NO}_{2}^{-}\right)$. If this would be the only process in operation, all freed nitrogen gas would be of type ${ }^{28} N_{2}$, as indicated by the green dot in Figure 1. Alternatively, nitrite can be produced by nitrate reduction $\left({ }^{14,15} \mathrm{NO}_{3}^{-} \longrightarrow \longrightarrow^{14,15} \mathrm{NO}_{2}^{-}\right)$. If the latter would be the only process in operation, no ${ }^{30} N_{2}$ would be produced, and the final nitrogen gas composition is restricted to be on an edge of the ternary plot, as indicated by the red line in Figure 1.

The combination of these processes determines the final composition of the freed nitrogen gas, and it is of interest to quantify the relative contribution of these processes to the final composition.


Figure 1: Ternary plot for nitrogen compositions. The blue curve corresponds to classical denitrification (HardyWeinberg equilibrium). The red dot corresponds to a sample with proportions $0.3,0.6$ and 0.1 of ${ }^{28} N_{2},{ }^{29} N_{2}$ and ${ }^{30} N_{2}$ respectively, and this composition has a fraction of 0.4 of isotope ${ }^{15} N$ and a fraction of 0.6 of isotope ${ }^{14} N$. The green dot represents Anammox combined with ammonium oxidation. The red edge represents Anammox with nitrate reduction.

We explore the effect of mixing the compositions obtained by classical denitrification and the Anammox process by combining the theoretical compositions under classical denitrification and Anammox in combination with nitrate reduction by taking a convex linear combination of the composition of two processes. That is to say, the finally resulting observed composition $\mathbf{y}=$ $\left({ }^{28} N_{2},{ }^{29} N_{2},{ }^{30} N_{2}\right)$ is thought to originate as

$$
\mathbf{y}=t \cdot \mathbf{x}_{c d}+(1-t) \cdot \mathbf{x}_{a n}
$$

where $\mathbf{x}_{c d}$ represents the nitrogen composition produced by classical denitrification $(c d), \mathbf{x}_{a n}$ the
nitrogen composition produced by Anammox with nitrate reduction (an), and $t$ a mixing proportion $(0 \leq t \leq 1)$. Though $\mathbf{x}_{c d}$ and $\mathbf{x}_{a n}$ are not experimentally directly measured, they could be estimated as $\left((1-p)^{2}, 2 p(1-p), p^{2}\right)$ and $(1-p, p, 0)$ respectively. Likewise, the observed compositions can also be conceived as a mixing of classical denitrification ( $c d$ ) and Anammox with ammonium oxidation (aa) giving

$$
\mathbf{y}=t \mathbf{x}_{c d}+(1-t) \mathbf{x}_{a a}
$$

where $\mathbf{x}_{a a}$ would be represented by $(1,0,0)$. The effect of mixing the two processes on the final composition is graphed in the ternary diagrams in Figure 2 for different values of $t$.


Figure 2: Nitrogen compositions obtained under mixing of classical denitrification with Anammox and nitrate reduction (A) and mixing of classical denitrification with Anammox and ammonium oxidation (B).

Figure 2A shows that classical denitrification and Anammox in combination with nitrate reduction gives compositional curves that are always above the Hardy-Weinberg parabola. The HardyWeinberg parabola is obtained for $t=1$. Classical denitrification and Anammox with ammonium oxidation gives parabolas that are always below the Hardy-Weinberg parabola (Fig. 2B). In its most general formulation, the observed nitrogen composition is the result of the three aforementioned processes, and a model for the data could be written as

$$
\begin{equation*}
\mathbf{y}=t_{1} \cdot \mathbf{x}_{c d}+t_{2} \cdot \mathbf{x}_{a n}+t_{3} \cdot \mathbf{x}_{a a}+\varepsilon \tag{1}
\end{equation*}
$$

with mixing constants satisfying $t_{1}+t_{2}+t_{3}=1$ and $0 \leq t_{i} \leq 1$, and $\varepsilon$ an error term. It of ecological interest to estimate the relative contribution of the three processes, i.e. to estimate the mixing parameters in Equation (1). This is part of ongoing work.

## 2 Experimental data and results

A set of 11 sediment samples (cores sampled in Etang du Méjean close to Montpellier in France) was incubated with different relative amounts of ${ }^{15} \mathrm{NO}_{3}^{-}$added in the water column (fractions of 0.00 , $0.25,0.50,0.75$ or 1.00 of ${ }^{15} \mathrm{NO}_{3}^{-}$with respect to the total amount of nitrate $\left({ }^{15} \mathrm{NO}_{3}^{-}+{ }^{14} \mathrm{NO}_{3}^{-}\right)$), and the relative amounts of ${ }^{28} N_{2},{ }^{29} N_{2}$ and ${ }^{30} N_{2}$ produced were measured by mass-spectrometry (Kana and others, 1998). The experimental data are represented in Figure 3 and are seen to fall
on or below the Hardy-Weinberg parabola of classical denitrification. The observed proportions of ${ }^{15} \mathrm{~N}$ in nitrogen gas are seen to be somewhat below the experimentally applied proportions of ${ }^{15} \mathrm{NO}_{3}^{-}$in the water column. This can be explained by the fact that some ${ }^{14} \mathrm{NO}_{3}^{-}$was present in the sediments, where it was most likely produced through nitrification. Hence, the nitrate diffusing from the water column $\left({ }^{15} \mathrm{NO}_{3}^{-} /{ }^{14} \mathrm{NO}_{3}^{-}\right.$according the experimental addition in the water column) mixes in the sediment with this ${ }^{14} \mathrm{NO}_{3}^{-}$to result in a lower ${ }^{15} \mathrm{NO}_{3}^{-} /{ }^{14} \mathrm{NO}_{3}^{-}$in the sedimentary reaction zone where the $N_{2}$ is produced. It is of interest to formally test the null hypothesis of 100


Figure 3: Ternary plot for nitrogen compositions of 11 sediment samples incubated with different amounts of the ${ }^{15} N$ isotope.
percent classical denitrification. This can be done by a test on the mean of the isometric log-ratio coordinates, as proposed by Graffelman \& Egozcue (2011). Under the null hypothesis of HardyWeinberg equilibrium, the mean of the second isometric $\log$-ratio coordinate is $-\sqrt{2 / 3} \ln (2)$. The null of $100 \%$ classical denitrification can thus be tested by a one-sample $t$-test on the second isometric $\log$-ratio coordinate. This gives $T=3.99$ with a $p$-value 0.0025 , leading us to reject the null hypothesis. This suggests that classical denitrification, although predominant, coexisted with Anammox coupled to ammonium oxidizers. It is clear that the pattern of the observed compositions in Figure 3 can not be explained by a combination of classical denitrification and Anammox coupled to nitrate reduction alone.

## 3 Conclusions

Experimental data collected in microbial ecology like the data set studied here is sometimes of compositional nature. Graphical tools from compositional data analysis like ternary diagrams and
log-ratio transformations are useful for analyzing such datasets.

## Acknowledgements

This study was supported by grant CODARSS MTM2012-33236 of the Spanish Ministry of Education and Science. Emilie Le'Floch and Carla Pérez are thanked for their contribution to the experimental work. Support from the CNRS (INEE) and from the DEPART project (Agence de l'Eau - Rhône Méditerranée Corse, Conseil Régional du Languedoc-Roussillon) allowed to acquire the Membrane Inlet Mass Spectrometry equipment used in this study.

## References

Aitchison, J. (2003). The Statistical Analysis of Compositional Data. The Blackburn Press,Caldwell, NJ (USA). 435 p.

Deutsch, B., Forster, S., Wilhelm, M., Dippner, J. W. and Voss, M. (2010). Denitrification in sediments as a major nitrogen sink in the Baltic Sea: an extrapolation using sediment characteristics. Biogeosciences 7, pp. 3259--3271. doi:10.5194/bg-7-3259-2010

Graffelman, J. and Egozcue, J. J. (2011). Hardy-Weinberg equilibrium: a nonparametric compositional approach. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), Compositional Data Analysis: Theory and Applications, pp. 208-217, John Wiley \& Sons, Ltd.

Graffelman, J. and Morales-Camarena, J. (2008). Graphical tests for Hardy-Weinberg equilibrium based on the ternary plot. Human Heredity 65(2), pp. 77-84.

Kana, T. M., Sullivan, M. B., Cornwell, J.C. and Groszkowski, K. (1998). Denitrification in estuarine sediments determined by membrane inlet mass spectrometry. Limnology and Oceanography 43, pp. 334-339.

Kuenen, J. G. (2008). Anammox bacteria: from discovery to application. Nature Reviews Microbiology 6, pp. 320-326.

Nielsen, L. P. (1992). Denitrification in sediment determined from nitrogen isotope pairing. FEMS Microbiology Ecology 86, pp. 357-362.

# On the asymptotic distribution of proportions of multinomial count data 

J. Graffelman ${ }^{1}$, J.J. Egozcue ${ }^{1}$, and M.I. Ortego ${ }^{1}$<br>${ }^{1}$ Universitat Politècnica de Catalunya, Barcelona, Spain; jan.graffelman@upc.edu


#### Abstract

Compositional data is often transformed by taking log-ratios prior to the statistical analysis. The transformed variables are usually assumed to be approximately normally distributed. In this situation, random compositions follow (approximately) a normal distribution on the simplex. This justifies the posterior use of standard statistical techniques that rely on the normality assumption, such as Student $t$-tests, regression modeling, analysis of variance and other techniques and models. As a consequence, results supporting the asymptotic normality of log-ratio transformations are worth to be studied. In this contribution, a multinomial sampling scheme is considered. The distribution of the proportions of the counts in each category are transformed by an isometric log-ratio transformation (ilr). The asymptotic distribution of these coordinates is shown to be normal, and its mean and covariance parameters are expressed as functions of the multinomial probabilities. The theory is based on the known mean and variance of a multinomial variable and its asymptotic normal distribution. Following a previous study for trinomial variables (Graffelman, 2011), the delta-method, applied to the maximum likelihood estimators of frequencies, provides the asymptotic normal distribution of the isometric log-ratio coordinates. As a consequence, proportions from multinomial counts are asymptotically normal on the simplex and its parameters (center and variability) depend only on the multinomial parameters. Theoretical and simulated examples are presented.


Key words: Delta-method, multinomial counts, compositional data, normal on the simplex.

## 1 Introduction

Multinomial sampling is a frequent way of collecting data. A random sample of individuals is classified into a set of well-defined categories. The final result is a vector of counts which reports how many individuals are classified in each category. Examples of this kind of sampling appear everywhere. In social science, surveyed individuals are classified according preferences, social classes, educational levels, age, gender, etc. In geoscience, millions of counts of different chemical elements are obtained from automatic spectrometers or other automatic measurement devices. Similarly, in biology and health science data may come from few counts of individuals or enormous amount of counts coming from genomic experiments. Most of the times, interest is not focused on the total number of counts but on the probability of the outcome of the multinomial experiment in a particular class. Accordingly, the relevant information to estimate these probabilities is conveyed by the proportions of counts assigned to each category or class. This very old but effective practice may become problematic when the number of counts in some categories is small or even zero. In this situation of scarcity of counts, the traditional way of estimation of the underlying probabilities, that is, the ratio of counts in a category over the total number of counts, becomes poor and confusing. A typical situation is that a category obtains a zero count in the sampling process, leading to an unreliable estimate (null) for that category. This problem affects not only zero-counts but all relatively small counts that typically occur for less probable categories. Dramatically, estimating these small probabilities is often the goal of the analysis.

In many scenarios, multinomial count data are presented as proportions. Assume that $n$ multinomial experiments provide counts classified into $D$ categories. The counts in such categories are denoted $\left(n_{1}, n_{2}, \ldots, n_{D}\right)$, where $n_{i}$ are non-negative integers. The proportions of counts or relative frequencies are then $\mathbf{f}=\left(f_{1}, f_{2}, \ldots, f_{D}\right)$, with $f_{i}=n_{i} / n$, which are the maximum likelihood estimators of the multinomial probabilities $\mathbf{p}=\left(p_{1}, p_{2}, \ldots, p_{D}\right)$. The asymptotic distribution of $\mathbf{f}$ is known to be a multivariate normal distribution when considering $\mathbf{f}$ as a real vector of $\mathbb{R}^{D}$. Historically, the first result of this kind was the DeMoivre-Laplace theorem (Chow and Teicher, 1997). Extensions for the multinomial distribution are available elsewhere (Rohatgi, 1976). The use of the multivariate normal distribution, as an asymptotic approximation of the multinomial distribution or its respective frequencies, is in contradiction with the nature of the vector of proportions $\mathbf{f}$. The multivariate normal distribution is supported on the whole $\mathbb{R}^{D}$, while the vector of proportions $\mathbf{f}$ can only take values in a unitary $D$-part simplex, $\mathcal{S}^{D}$, which is a restricted manifold within $\mathbb{R}^{D}$. With small or moderate counts, the asymptotic multivariate normal distribution assigns some non-zero probability to events of negative proportions or proportions larger than one and thus becomes an inappropriate model.

The practice in compositional data analysis consists of assuming that the multinomial parameters p constitute a composition in $\mathcal{S}^{D}$ (Pawlowsky-Glahn et al., 2015). Accordingly, estimators of p such as $\mathbf{f}$ are also assumed compositional. The simplex $\mathcal{S}^{D}$ endowed with its Aitchison geometry is a Euclidean vector space (Pawlowsky-Glahn and Egozcue, 2001). As a consequence, its elements, such as $\mathbf{p}$ and $\mathbf{f}$, can be represented by Cartesian coordinates, commonly called ilr-coordinates (Egozcue et al., 2003).

The most important probability distribution defined in the simplex is the logistic normal distribution (Aitchison and Shen, 1980; Aitchison, 1982), also known as normal on the simplex (MateuFigueras et al., 2013). Its relevance comes from the fact that the normal distribution on the simplex is the distribution appearing as a central limit distribution when a large number of random compositions is repeatedly perturbed (Aitchison, 1986). A random composition is said to be normally distributed on the simplex when its random ilr-coordinates follow a multivariate normal distribution. Due to these properties, the normal distribution on the simplex is the first candidate to be an asymptotic distribution in many situations and, particularly, in the present case.

The main goal of this study is to investigate the probability distribution of the ilr coordinates of the estimated frequencies $\mathbf{f}$, under multinomial sampling. The multivariate normal distribution has been used for the log-ratio coordinates of $\mathbf{f}$ for several decades, but a formal proof for the
asymptotic normality of the ilr-coordinates seems not to have been provided. The expectation of the coordinates was intuitively and correctly proposed, but the description of their variability has remained unclear. This contribution provides a solution for this problem. The remainder of this paper is organized as follows. Section 2 recalls some known concepts regarding ilr-coordinates and the Aitchison geometry of the simplex. Section 3 develops the asymptotic distribution theory using the so-called delta-method (Casella and Berger, 2002). Finally, Section 4 presents some examples where the asymptotic distribution plays a role.

## 2 Coordinates and normal distribution on the simplex

Let $\mathbf{x}=\left(x_{1}, x_{2}, \ldots, x_{D}\right)$ be a composition in $\mathcal{S}^{D}$, which will be treated as a column vector when using matrix notation. It is assumed that $\sum_{i=1}^{D} x_{i}=1$, although vectors with proportional positive components are compositionally equivalent. For a vector $\mathbf{w}$ with positive components which do not add to 1 , define the closure operation $\mathcal{C}(\mathbf{w})$ consisting of dividing each component by the sum of them, so that the resulting components add to 1 .

As mentioned, the simplex $\mathcal{S}^{D}$, endowed with the perturbation and powering operations, and the Aitchison inner product, is a Euclidean space of dimension $D-1$. Cartesian coordinates of $\mathbf{x}$, called ilr-coordinates are obtained through a so called contrast matrix $\mathbf{V}$, which is of order $D$ by $D-1$ and satisfies the following properties

$$
\mathbf{V}^{\prime} \mathbf{V}=\mathbf{I}_{(D-1)} \quad, \quad \mathbf{V} \mathbf{V}^{\prime}=\mathbf{I}_{D}-\frac{1}{D} \mathbf{1 1}^{\prime}
$$

where $\mathbf{I}_{k}$ is the identity matrix of order $k$ by $k$ and $\mathbf{1}$ is a column vector with ones in all its components (Egozcue et al., 2011). The $D-1$ ilr-coordinates of $\mathbf{x}$, associated with a given contrast matrix $\mathbf{V}$, are obtained as

$$
\begin{equation*}
\operatorname{ilr}(\mathbf{x})=\mathbf{y}=\mathbf{V}^{\prime} \ln (\mathbf{x}) \quad, \quad \mathbf{x}=\operatorname{ilr}^{-1}(\mathbf{y})=\mathcal{C}(\exp (\mathbf{V} \mathbf{y})) \tag{1}
\end{equation*}
$$

where $\ln$ and exp operate componentwise.
A comfortable and intuitive way of obtaining $\mathbf{V}$ and the corresponding ilr-coordinates is by defining a sequential binary partition (SBP) (Egozcue and Pawlowsky-Glahn, 2005, 2006). It consist of partitioning the parts of the composition $\mathbf{x}$ into two groups of parts, coded as +1 and -1 . Each of these groups are divided into two groups again, until all remaining groups contain only one part. When a part does not participate in the partition of a group it is marked with a zero. The following array of codes is an example for a 5-part composition; each column represents a partition.

$$
\boldsymbol{\Theta}=\left[\begin{array}{rrrr}
+1 & +1 & 0 & 0  \tag{2}\\
+1 & -1 & 0 & +1 \\
+1 & -1 & 0 & -1 \\
-1 & 0 & +1 & 0 \\
-1 & 0 & -1 & 0
\end{array}\right]
$$

From this kind of SBP code matrices, the corresponding contrast matrix is built from

$$
v_{i j}=\theta_{i j} \frac{1}{n_{j}^{\left(\theta_{i j}\right)}} \sqrt{\frac{n_{j}^{+} \cdot n_{j}^{-}}{n_{j}^{+}+n_{j}^{-}}}, i=1,2, \ldots, D ; j=1,2, \ldots, D-1
$$

where $n_{j}^{+}, n_{j}^{-}$are the number of $+1,-1$ in the $j$-column respectively, and $n_{j}^{\theta_{i j}}$ is one of $n_{j}^{+}, n_{j}^{-}$ depending of the sign in $\theta_{i j}$. For instance, the first column of the contrast matrix $\mathbf{V}$ corresponding to the code $\boldsymbol{\Theta}$ in (2) is

$$
\mathbf{v}_{1}=\left(+\frac{1}{3} \sqrt{\frac{6}{5}},+\frac{1}{3} \sqrt{\frac{6}{5}},+\frac{1}{3} \sqrt{\frac{6}{5}},-\frac{1}{2} \sqrt{\frac{6}{5}},-\frac{1}{2} \sqrt{\frac{6}{5}}\right)^{\prime}
$$

The expression of ilr-coordinates corresponding to the contrast matrix $\mathbf{X}$ are obtained using Equation (1). The $j$-coordinate is

$$
y_{j}=\operatorname{ilr}_{j}(\mathbf{x})=\sqrt{\frac{n_{j}^{+} \cdot n_{j}^{-}}{n_{j}^{+}+n_{j}^{-}}} \ln \frac{g\left(\mathbf{x}_{j}^{+}\right)}{g\left(\mathbf{x}_{j}^{-}\right)},
$$

where $g\left(\mathbf{x}_{j}^{+}\right)$and $g\left(\mathbf{x}_{j}^{-}\right)$are the geometric mean of parts of $\mathbf{x}$ coded as $+1,-1$ in the $j$-column of $\boldsymbol{\Theta}$. These kind of coordinates are called balances for the ratio of geometric means appearing as argument of the logarithm which, depending on the problem, may be easily interpretable. For instance, the first balance-coordinate associated with the column $\mathbf{v}_{1}$ is

$$
y_{1}=\sqrt{\frac{6}{5}} \ln \frac{\left(x_{1} x_{2} x_{3}\right)^{1 / 3}}{\left(x_{4} x_{5}\right)^{1 / 2}} .
$$

When $\mathbf{x}$ is a random composition, its ilr-coordinates are also random and the distribution of $\mathbf{x}$ can be defined through the distribution of its ilr-coordinates. Particularly, $\mathbf{x}$ has a normal distribution on the simplex whenever its ilr-coordinates are multivariate normal distributed. The parameters of the normal on the simplex are usually taken as the mean, $\boldsymbol{\mu}_{y}$, and variance-covariance matrix $\boldsymbol{\Sigma}_{y}$ of the ilr-coordinates. These facts can be written as

$$
\mathbf{x} \sim \mathcal{N}_{\mathcal{S}^{D}}\left(\boldsymbol{\mu}_{y}, \boldsymbol{\Sigma}_{y}\right) \quad, \quad \mathbf{y} \sim \mathcal{N}_{D-1}\left(\boldsymbol{\mu}_{y}, \boldsymbol{\Sigma}_{y}\right)
$$

where it should be taken into account that parameters ( $\boldsymbol{\mu}_{y}, \boldsymbol{\Sigma}_{y}$ ) depend on the selected contrast matrix for performing the ilr-coordinates (Pawlowsky-Glahn et al., 2015).

## 3 Asymptotic theory

Consider a multinomial sampling scenario in which $n$ multinomial observations are given, and the observed relative frequencies are in $\mathbf{f}$. As $\mathbf{f}$ is the maximum likelihood estimator of the the multinomial probabilities $\mathbf{p}$, in asymptotic conditions, the distribution of $\mathbf{f}$ approaches (weak convergence in distribution or law) to the multivariate normal distribution

$$
\mathbf{f} \xrightarrow{D} \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{f}}, \boldsymbol{\Sigma}_{\boldsymbol{f}}\right) \quad, \quad \boldsymbol{\mu}_{f}=\mathbf{p} \quad, \quad \boldsymbol{\Sigma}_{f}=\mathbf{D}_{p}-\mathbf{p p}^{\prime}
$$

where $\mathbf{D}_{p}$ is diagonal and has the multinomial probabilities in $\mathbf{p}$ as entries. The present goal is to look for the asymptotic distribution of the ilr-coordinates of $\mathbf{f}$, denoted $\hat{\mathbf{y}}$, associated with a given contrast matrix $\mathbf{V}$, as was developed for $D=3$ by Graffelman (2011).
For a regular transformation of parameters $\mathbf{g}$, according to the multivariate delta method (Casella and Berger, 2002), the asymptotic centered distribution of the maximum likelihood estimator $\mathbf{g}(\mathbf{f})$ is

$$
\begin{equation*}
\sqrt{n}(\mathbf{g}(\mathbf{f})-\mathbf{g}(\mathbf{p})) \approx \mathcal{N}\left(\mathbf{0},\left(\frac{\partial \mathbf{g}}{\partial \mathbf{p}}\right) \boldsymbol{\Sigma}_{f}\left(\frac{\partial \mathbf{g}}{\partial \mathbf{p}}\right)^{\prime}\right) \tag{3}
\end{equation*}
$$

where $\boldsymbol{\Sigma}_{f}$ is the covariance matrix of $\mathbf{f}$. When applied to the maximum likelihood estimator $\mathbf{f}$ of the multinomial probabilities $\mathbf{p}$, transformed into $\mathbf{y}=\operatorname{ilr}(\mathbf{p})$, the transformation is identified as $\mathbf{g}(\mathbf{p})=\operatorname{ilr}(\mathbf{p})$ which is given in Equation (1). The computation of derivatives in Equation (3) can be carried out as

$$
\frac{\partial \operatorname{iilr}(\mathbf{p})}{\partial \mathbf{p}}=\frac{\partial \mathbf{V}^{\prime} \ln \mathbf{p}}{\partial \mathbf{p}}=\mathbf{V}^{\prime}\left[\begin{array}{cccc}
\frac{1}{p_{1}} & & & \\
& \frac{1}{p_{2}} & & \\
& & \ddots & \\
& & & \frac{1}{p_{D}}
\end{array}\right]=\mathbf{V}^{\prime} \mathbf{D}_{p}^{-1}
$$

The covariance matrix $\boldsymbol{\Sigma}_{\hat{y}}$ becomes

$$
\boldsymbol{\Sigma}_{\hat{y}}=\mathbf{V}^{\prime} \mathbf{D}_{p}^{-1} \boldsymbol{\Sigma}_{f} \mathbf{D}_{p}^{-1} \mathbf{V}=\mathbf{V}^{\prime} \mathbf{D}_{p}^{-1}\left(\mathbf{D}_{p}-\mathbf{p p}^{\prime}\right) \mathbf{D}_{p}^{-1} \mathbf{V}=\mathbf{V}^{\prime} \mathbf{D}_{p}^{-1} \mathbf{V}-\mathbf{V}^{\prime} \mathbf{1} 1^{\prime} \mathbf{V}=\mathbf{V}^{\prime} \mathbf{D}_{p}^{-1} \mathbf{V}
$$

since $\mathbf{V}^{\prime} \mathbf{1}=\mathbf{0}$. Note that the covariance matrix $\mathbf{V}^{\prime} \mathbf{D}_{p}^{-1} \mathbf{V}$ is, in general, not diagonal and, therefore, the ilr-coordinates of $\mathbf{f}$ will be correlated. However, these correlations have a structure that only depends on $\mathbf{p}$ and the chosen contrast matrix $\mathbf{V}$. The matrix $\mathbf{V}^{\prime} \mathbf{D}_{p}^{-1} \mathbf{V}$ is diagonal only for equal multinomial probabilities, which correspond to $\mathbf{p}$ equal to the neutral element of the simplex. We note that the total variance in $\boldsymbol{\Sigma}_{\hat{y}}$ is given by

$$
\operatorname{tr}\left(\mathbf{V}^{\prime} \mathbf{D}_{p}^{-1} \mathbf{V}\right)=\operatorname{tr}\left(\mathbf{V} \mathbf{V}^{\prime} \mathbf{D}_{p}^{-1}\right)=\operatorname{tr}\left(\left(\mathbf{I}-\frac{1}{D} \mathbf{1 1}\right) \mathbf{D}_{p}^{-1}\right)=\operatorname{tr}\left(\mathbf{D}_{p}^{-1}\right)-\operatorname{tr}\left(\frac{1}{D} \mathbf{1 1}^{\prime} \mathbf{D}_{p}^{-1}\right)=\frac{D-1}{D} \sum_{i=1}^{D} \frac{1}{p_{i}}
$$

which effectively only depends on $\mathbf{p}$. Substituting the values of derivatives and the covariance matrix in Equation (3) the asymptotic distribution of $\hat{\mathbf{y}}=\operatorname{ilr}(\mathbf{f})$ satisfies

$$
\sqrt{n}(\hat{\mathbf{y}}-\mathbf{y}) \xrightarrow{D} \mathcal{N}_{D-1}\left(\mathbf{0}, \mathbf{V}^{\prime} \mathbf{D}_{p}^{-1} \mathbf{V}\right)
$$

where $\mathbf{y}=\operatorname{ilr}(\mathbf{p})$; the $D$ over the limit arrow denotes convergence in distribution or in law. This statement can be alternatively written as

$$
\begin{equation*}
\hat{\mathbf{y}} \xrightarrow{D} \mathcal{N}_{D-1}\left(\mathbf{y}, \frac{1}{n} \mathbf{V}^{\prime} \mathbf{D}_{p}^{-1} \mathbf{V}\right) \tag{4}
\end{equation*}
$$

## 4 Examples

We present two examples of our asymptotic result in Equation (4). The first example is theoretical, and obtains the covariance matrix of the ilr-coordinates for a genetic problem. The second example is numerical and investigates the convergence to the asymptotic distribution.

### 4.1 Compositional test for Hardy-Weinberg equilibrium for a diallelic genetic marker

Consider the following sequential binary partition for the three genotypes $A A, B B, A B$ with respective frequencies $p_{A A}, p_{B B}$ and $p_{A B}$ :

| $A A$ | $B B$ | $A B$ |
| ---: | ---: | ---: |
| +1 | +1 | -1 |
| +1 | -1 | 0 |

The first partition is a balance between homozygotes and heterozygotes, and the second partition splits the $A A$ and $B B$ homozygote. The associated contrast matrix is

$$
\mathbf{V}=\left[\begin{array}{rr}
+1 / \sqrt{6} & +1 / \sqrt{2}  \tag{5}\\
+1 / \sqrt{6} & -1 / \sqrt{2} \\
-2 / \sqrt{6} & 0
\end{array}\right]
$$

Then, the asymptotic covariance matrix of the ilr-coordinates of $\mathbf{f}$ is

$$
\begin{aligned}
& \mathbf{V}^{\prime} \mathbf{D}_{p}^{-1} \mathbf{V}=\left[\begin{array}{ccc}
\frac{+1}{\sqrt{6}} & \frac{+1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} \\
\frac{+1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0
\end{array}\right]\left[\begin{array}{ccc}
\frac{1}{p_{A A}} & 0 & 0 \\
0 & \frac{1}{p_{B B}} & 0 \\
0 & 0 & \frac{1}{p_{A B}}
\end{array}\right]\left[\begin{array}{cc}
\frac{+1}{\sqrt{6}} & \frac{+1}{\sqrt{2}} \\
\frac{+1}{\sqrt{6}} & \frac{-1}{\sqrt{2}} \\
\frac{-2}{\sqrt{6}} & 0
\end{array}\right] \\
&=\left[\begin{array}{cc}
\frac{1}{6}\left(\frac{1}{p_{A A}}+\frac{4}{p_{A B}}+\frac{1}{p_{B B}}\right) & \frac{1}{2 \sqrt{3}}\left(\frac{1}{p_{A A}}-\frac{1}{p_{B B}}\right) \\
\frac{1}{2 \sqrt{3}}\left(\frac{1}{p_{A A}}-\frac{1}{p_{B B}}\right) & \frac{1}{2}\left(\frac{1}{p_{A A}}+\frac{1}{p_{B B}}\right)
\end{array}\right],
\end{aligned}
$$

which is exactly the covariance matrix obtained previously in Graffelman (2011).


Figure 1: Difference between the sample mean value of balances and asymptotic mean value. First, second, third and fourth balances in black, blue, red and green curves respectively.

### 4.2 Convergence to the asymptotic distribution

A simulation of multinomial sampling has been conducted in order to check convergence of balancecoordinates to the asymptotic parameters of the distribution (4). A multinomial experiment with probabilities $\mathbf{p}=(0.03,0.20,0.30,0.40,0.07)$ has been chosen. The size of the samples examined ranges from $n=5$ to $n=600$ increasing by 5 . For each sample size $n$, 100 realizations of the multinomial distribution have been drawn and then transformed into ilr-balances using the contrast matrix $\mathbf{V}$ corresponding to the sign code shown in Equation (2).
According to Equation (4), the asymptotic distribution of the maximum likelihood estimators of $\operatorname{ilr}(\mathbf{p})$, denoted here as $\hat{\mathbf{y}}$, is a multivariate normal distribution with $\mathrm{E}[\hat{\mathbf{y}}]=\operatorname{ilr}(\mathbf{p})$ and (rounded) covariance matrix

$$
\boldsymbol{\Sigma}_{\hat{\boldsymbol{y}}}=\frac{1}{n} \mathbf{V}^{\prime} \mathbf{D}_{p}^{-1} \mathbf{V}=\frac{1}{n}\left(\begin{array}{rrrr}
10.59 & 8.70 & 4.56 & 0.43 \\
8.70 & 23.61 & 0.00 & -0.48 \\
4.56 & 0.00 & 8.39 & 0.00 \\
0.43 & -0.48 & 0.00 & 4.17
\end{array}\right)
$$

with eigenvalues $(1 / n)(28.19,11.40,4.31,2.87)$. For each value of $n$, the 100 multinomial realizations have been transformed into ilr-coordinates (see Eq. (1)). The sample mean vector and sample covariance matrix of these 100 -samples is then computed; also the eigenvalues of the sample covariance matrix are computed. Figure 1 shows the difference between the estimated mean balance and the given $\operatorname{ilr}(\mathbf{p})$. Although convergence to the true values is rather quick with $n$, it should be pointed out that for $n<100$ the approximation is hardly reliable. From simulations, the sample covariances of the balance coordinates have been computed, and then, compared with the asymptotic values using the corresponding eigenvalues. Figure 2 (left panel) shows the log-ratios of the eigenvalues of the sample covariance matrix over the corresponding asymptotic one. A null value indicates perfect match. Eigenvalues are ordered in decreasing magnitude and the curves are colored as black (largest eigenvalue), blue, red and green (smallest eigenvalue). The convergence rate seems not depend on the magnitude of the eigenvalue.

An alternative approach for estimation of the covariance matrix is provided by the asymptotic distribution in Equation (4). In fact, the covariance matrix of the balance-coordinates can be estimated by replacing $\mathbf{p}$ in Equation (4) with the back-transformed ( $\mathrm{ilr}^{-1}$ ) mean of the simulated balances. This gives a better convergence as shown in Figure 2 (right panel). The left panel clearly shows a larger variability when sample ilr covariances are used to estimate $\boldsymbol{\Sigma}_{\hat{y}}$. In Figure 2 (right panel) the largest eigenvalue converges slower than other ones; however, this was not confirmed by repeated simulations.


Figure 2: Comparison of the sample eigenvalues of the covariance matrix of simulated balances. Curves are the log-ratio of each sample eigenvalue over the asymptotic one; perfect approximation correspond to a zero value. First, second, third and fourth eigenvalues are the black, blue, red and green curves respectively. Left panel: sample covariances are directly estimated on simulated balances; Right panel: covariances are estimated from the sample means of balances.

## 5 Conclusion

The main result is the explicit form of the asymptotic distribution of the ilr-coordinates of a multinomial sample. When the total number of counts of the multinomial sample is $n$, the distribution converges in law to a multivariate normal distribution, whose mean values are the multinomial probabilities and the covariance matrix is $\frac{1}{n} \mathbf{V}^{\prime} \mathbf{D}_{p}^{-1} \mathbf{V}$. Consequently, the distribution of the proportions of multinomial counts are asymptotically normally distributed on the simplex.

This result shows the normal on the simplex to be the first choice for modelling the distribution of proportions of counts, and also shows the relationship between the centre of the distribution and its variability.

## Acknowledgements

This research has been supported by the Spanish Ministry of Education and Science under project 'METRICS' (Ref. MTM2012-33236); and by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project Ref. 2009SGR424.

## REFERENCES

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). Journal of the Royal Statistical Society, Series B (Statistical Methodology) 44 (2), 139-177.

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman \& Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Aitchison, J. and S. M. Shen (1980). Logistic-normal distributions. Some properties and uses. Biometrika 67(2), 261-272.

Casella, R. and R. L. Berger (2002). Statistical Inference. Second edition, Duxbury, Pacific Grove, CA, USA.

Chow, Y. S. and H. Teicher (1997). Probability Theory: Independence, Interchangeability, Martingales. Springer Texts in Statistics. Springer-Verlag, New York, NY (USA). 488 p.

Egozcue, J. J., C. Barceló-Vidal, J. A. Martín-Fernández, E. Jarauta-Bragulat, J.-L. DíazBarrero, and G. Mateu-Figueras (2011). Elements of simplicial linear algebra and geometry. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), Compositional Data Analysis: Theory and Applications., pp. 141-157. John Wiley \& Sons. 378 p.

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

Egozcue, J. J. and V. Pawlowsky-Glahn (2006). Simplicial geometry for compositional data. In: Buccianti, A., Mateu-Figueras, G. and Pawlowsky-Glahn, V., (eds) Compositional Data Analysis in the Geosciences: From Theory to Practice. Geological Society, London, Special Publications, 264, 67-77.

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

Graffelman, J. (2011). Statistical inference for Hardy-Weinberg equilibrium using log-ratio coordinates. In J. J. Egozcue, R. Tolosana-Delgado, and M. I. Ortego (Eds.), Proceedings of the 4 th International Workshop on Compositional Data Analysis (2011), Sant Feliu de Guíxols, Girona, Spain, May 9-13. CIMNE, Barcelona, Spain ISBN 978-84-87867-76-7.

Mateu-Figueras, G., V. Pawlowsky-Glahn, and J. J. Egozcue (2013). The normal distribution in some constrained sample spaces. SORT - Statistics and Operations Research Transactions 37(1), 29-56.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment (SERRA) 15(5), 384-398.

Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2015). Modeling and analysis of compositional data. Statistics in practice. John Wiley \& Sons, Chichester UK. 272 pp.

Rohatgi, V. K. (1976). An introduction to probability theory and mathematical statistics. Wiley Series in Probability and Statistics. John Wiley and Sons, New York, NY (USA). 684 p.

# An application of the isometric log-ratio transformation in relatedness research 

J. Graffelman ${ }^{1}$ and I. Galván ${ }^{2}$<br>${ }^{1}$ Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain; jan.graffelman@upc.edu<br>${ }^{2}$ Department of Computer Science, Applied Mathematics and Statistics, Universitat de Girona; Girona, Spain; ivan.galvan@udg.edu


#### Abstract

Genetic marker data contains information on the degree of relatedness of a pair of individuals. Relatedness investigations are usually based on the extent to which alleles of a pair of individuals match over a set of markers for which their genotype has been determined. A distinction is usually drawn between alleles that are identical-by-state (IBS) and alleles that are identical-by-descent (IBD). Since any pair of individuals can only share 0,1 or 2 alleles IBS or IBD for any marker, 3-way compositions can be computed that consist of the fractions of markers sharing 0,1 or 2 alleles IBS (or IBD) for each pair. For any given standard relationship (e.g. parent-offspring, sister-brother, etc.) the probabilities $k_{0}, k_{1}$ and $k_{2}$ of sharing 0,1 or 2 IBD alleles are easily deduced and are usually referred to as Cotterman's coefficients. Marker data can be used to estimate these coefficients by maximum likelihood (Thompson, 1975; 1991). This maximization problem has the 2 -simplex as its domain. If there is no inbreeding, then the maximum must occur in a subset of the 2 -simplex. The maximization problem is then subject to an additional non-linear constraint $\left(k_{1}^{2} \geq\right.$ $4 k_{0} k_{2}$ ). Special optimization routines are needed that do respect all constraints of the problem. A re-parametrization of the likelihood in terms of isometric log-ratio (ilr) coordinates greatly simplifies the maximization problem. In isometric log-ratio coordinates the domain turns out to be rectangular, and maximization can be carried out by standard general-purpose maximization routines. We illustrate this point with some examples using data from the HapMap project.


Key words: Genetic marker, allele, identity-by-state, identity-by-descent.

## 1 Introduction

Methods for investigating the degree of relatedness of a pair of individuals form an active area of research in statistical genetics. Relatedness studies are based on the idea that genetic markers carry information regarding the family relationships of the individuals involved. There are several reasons for performing a relatedness study. First of all, such studies serve to verify documented relationships between individuals. If sufficient genetic information is available, a relatedness study may reveal that a pair of putative sibs is in fact a pair of half-sibs or a twin pair. An important criterion in relatedness studies is the degree to which a pair of individuals shares alleles over a set of genetic markers. If two individuals share many alleles at many loci then it becomes more likely that they are closely related. In the most extreme case, if all alleles at all loci coincide for a pair of individuals, then the pair is, supposing sufficiently polymorphic loci, in theory a monozygotic twin pair. However, some caution is called for, because $100 \%$ coincidence will also arise if two registers in the database are accidentally duplicated, or if a biological sample has been genotyped twice in the laboratory.

Secondly, relatedness is a reason of concern in gene-disease association studies for statistical reasons. The presence of related individuals violates the independence assumption that underlies many statistical techniques used in these studies, such as chi-square tests on contingency tables, logistic regression and others. It is thus of interest to investigate the possible relatedness of the individuals in the sample prior to applying tests for association. If relatedness is detected, one individual of each related pair may be removed in order to maintain independence.

A distinction can be drawn between relatedness studies that are based on alleles that are identical-by-state (IBS) and alleles that are identical-by-descent (IBD). Two alleles are IBS if they are the same irrespective of their provenance, a situation that is statistically often referred to as a "match". In this contribution we restrict ourselves to IBD allele-sharing. Two alleles are IBD if they are IBS and have descended from the same parent. Because a child receives one allele from each parent, it shares one allele IBD with its father and one allele IBD with its mother. It is not possible for the child to receive zero or two IBD alleles from the father. The probabilities of sharing 0,1 or 2 IBD alleles for a given relationship are called Cotterman coefficients, and denoted by $k_{0}, k_{1}$ and $k_{2}$ respectively. If $X$ denotes the number of IBD alleles for a parent-offspring (PO) pair then its probability distribution is $k_{0}=P(X=0)=0, k_{1}=P(X=1)=1$ and $k_{2}=P(X=2)=0$. For other relationships it is only slightly more involved to obtain the theoretical IBD probabilities. Let $\alpha / \beta$ represent the paternal alleles and $A / B$ the maternal alleles of a couple sharing two alleles IBS. This couple can have four possible types of children $(\alpha / A),(\alpha / B),(\beta / A)$ and $(\beta / B)$, and the number of IBD alleles for each possible pair of sibs is shown in Table 1 below.

Table 1: Number of IBD alleles for all possible pairs of sibs descendant from a $(\alpha / \beta, A / B)$ couple.

|  | $\alpha / A$ | $\alpha / B$ | $\beta / A$ | $\beta / B$ |
| :---: | :---: | :---: | :---: | :---: |
| $\alpha / A$ | 2 | 1 | 1 | 0 |
| $\alpha / B$ | 1 | 2 | 0 | 1 |
| $\beta / A$ | 1 | 0 | 2 | 1 |
| $\beta / B$ | 0 | 1 | 1 | 2 |

From this table it is easily inferred that the IBD sharing probabilities for a pair of full sibs are given by $k_{0}=\frac{1}{4}, k_{1}=\frac{1}{2}$ and $k_{2}=\frac{1}{4}$. IBD probabilities can be estimated from the genotype data by maximum likelihood, as described in Section 2. If the estimated probabilities coincide with or are close to a set of theoretically known IBD probabilities for a given relationship, then that relationship is inferred.

The structure of the remainder of this paper is as follows. First we review the maximum likelihood estimation of IBD probabilities. Then we present a re-parametrization of the likelihood in terms of isometric log-ratios. We give some detailed examples with data from the HapMap project. Finally we discuss our results and provide some references.

## 2 Maximum likelihood estimation of IBD probabilities

Good accounts of the ML estimation of IBD probabilities are given by Thompson (1991) and Weir et al. (2006). We briefly review ML estimation here in order to provide a self-contained paper. The Cotterman coefficients can be obtained in a similar way as outlined in the introduction for all standard family relationships and are given in Table 2 below.

Table 2: IBD probabilities or Cotterman coefficients for some standard family relationships (MZ=Monozygotic twins, $\mathrm{PO}=$ Parent-offspring, $\mathrm{FS}=$ Full sibs, $\mathrm{HS}=$ Half sibs, $\mathrm{AV}=$ Avuncular, $\mathrm{GG}=$ Grandparent-grandchild, $\mathrm{FC}=\mathrm{First}$ cousins, $\mathrm{UN}=$ Unrelated).

| Relationship | $k_{0}$ | $k_{1}$ | $k_{2}$ |
| :--- | :---: | :---: | :---: |
| MZ | 0 | 0 | 1 |
| PO | 0 | 1 | 0 |
| FS | $\frac{1}{4}$ | $\frac{1}{2}$ | $\frac{1}{4}$ |
| HS | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 |
| AV | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 |
| GG | $\frac{1}{2}$ | $\frac{1}{2}$ | 0 |
| FC | $\frac{3}{4}$ | $\frac{1}{4}$ | 0 |
| UN | 1 | 0 | 0 |

Let $G_{1}$ and $G_{2}$ be the pair of genotypes observed at a locus for two individuals, and let $q(0,1$ or 2) represent the number of IBD alleles. By the law of total probability we have:

$$
\begin{equation*}
P\left(G_{1} \cap G_{2} \mid k_{0}, k_{1}, k_{2}\right)=P\left(G_{1} \cap G_{2} \mid q=0\right) k_{0}+P\left(G_{1} \cap G_{2} \mid q=1\right) k_{1}+P\left(G_{1} \cap G_{2} \mid q=2\right) k_{2} \tag{1}
\end{equation*}
$$

The probabilities $P\left(G_{1} \cap G_{2} \mid q=0\right)$ depend on the genotypes of the individuals and are calculated from the allele frequencies in the population. We denote different alleles by the letters $i, j, l$ and $m$ and let $p_{i}, p_{j}, p_{l}$ and $p_{m}$ be their corresponding allele frequencies. We denote genotypes by $i / j$, the slash separating the alleles found on the homologous chromosomes. E.g. if $G_{1}=i / i$ and $G_{2}=i / i$, then under the assumption of Hardy-Weinberg equilibrium we obtain:

$$
\begin{aligned}
& P\left(G_{1}=i / i \cap G_{2}=i / i \mid q=0\right)=P\left(G_{1}=i / i\right) P\left(G_{2}=i / i\right)=p_{i}^{2} p_{i}^{2}=p_{i}^{4}, \\
& P\left(G_{1}=i / i \cap G_{2}=i / i \mid q=1\right)=P\left(G_{1}=i / i\right) P\left(G_{2}=i / i\left|G_{1}=i / i\right| q=1\right)=p_{i}^{2} p_{i}=p_{i}^{3}, \\
& P\left(G_{1}=i / i \cap G_{2}=i / i \mid q=2\right)=P\left(G_{1}=i / i\right)=P\left(G_{2}=i / i\right)=p_{i}^{2}
\end{aligned}
$$

These probabilities are also determined for all other genotype pairs ( $(i / i, i / j),(i / i, j / j)$, etc. $)$ and the results are given in Table 3 below. If there are $n$ independent genetic markers, then the likelihood function for a pair of individuals can be written as:

$$
\begin{equation*}
L\left(k_{0}, k_{1}, k_{2} \mid G_{1} \cap G_{2}\right)=\prod_{i=1}^{n}\left(d_{0 i} k_{0}+d_{1 i} k_{1}+d_{2 i} k_{2}\right) \tag{2}
\end{equation*}
$$

where the coefficients $d_{0 i}, d_{1 i}$ and $d_{2 i}$ depend on the nature of the pair (possibilities given in Table 3) and on the allele frequencies of the corresponding markers. E.g. if for one marker both individuals are homozygous (an ( $i / i, i / i$ ) pair) then the contribution to the likelihood function is

Table 3: Probabilities of observing 0, 1 or 2 IBD alleles for all possible genotype pairs.

| Pair | Shared alleles | $q=0$ | $q=1$ | $q=2$ |
| :--- | :---: | :---: | :---: | :---: |
| $(i / i, i / i)$ | 2 | $p_{i}^{4}$ | $p_{i}^{3}$ | $p_{i}^{2}$ |
| $(i / i, j / j)$ | 0 | $p_{i}^{2} p_{j}^{2}$ | 0 | 0 |
| $(i / i, i / j)$ | 1 | $2 p_{i}^{3} p_{j}$ | $p_{i}^{2} p_{j}$ | 0 |
| $(i / i, j / m)$ | 0 | $2 p_{i}^{2} p_{j} p_{m}$ | 0 | 0 |
| $(i / j, i / j)$ | 2 | $4 p_{i}^{2} p_{j}^{2}$ | $p_{i} p_{j}\left(p_{i}+p_{j}\right)$ | $2 p_{i} p_{j}$ |
| $(i / j, i / m)$ | 1 | $4 p_{i}^{2} p_{j} p_{m}$ | $p_{i} p_{j} p_{m}$ | 0 |
| $(i / j, m / l)$ | 0 | $4 p_{i} p_{j} p_{m} p_{l}$ | 0 | 0 |

$p_{i}^{4} k_{0}+p_{i}^{3} k_{1}+p_{i}^{2} k_{2}$, with $p_{i}$ the $i$ th allele frequency of that marker. Taking logarithms, we search to maximize the log-likelihood function:

$$
\begin{equation*}
l\left(k_{0}, k_{1}, k_{2} \mid G_{1} \cap G_{2}\right)=\sum_{i=1}^{n} \ln \left(d_{0 i} k_{0}+d_{1 i} k_{1}+d_{2 i} k_{2}\right) \tag{3}
\end{equation*}
$$

where $k_{0}, k_{1}$ and $k_{2}$ are the parameters to be estimated, and the coefficients $d_{0 i}, d_{1 i}, d_{2 i}$ are obtained by substituting the sample allele frequencies in accordance with the type of pair.

## 3 Re-parametrization of the likelihood in isometric log-ratios

The maximization of the likelihood in Equation (3) is not trivial, as the following constraints need to be taken into account: $0 \leq k_{i} \leq 1, \sum_{i=0}^{2} k_{i}=1$ and $k_{1}^{2} \geq 4 k_{0} k_{2}$. The last inequality follows from the assumption of absence of inbreeding (Thompson, 1991). This is a maximization problem that has an arrow-headed subset of the simplex as its domain, as is shown by the grey region in Figure 1A. Standard $R$ functions like optim or nlminb from the stats package ( $R$ core team, 2014) assume the domain of the objective function to be rectangular. For these R functions the simplex constraint $k_{0}+k_{1}+k_{2}=1$ is a problem. The range for $k_{i}$ is not simply the [0,1] interval but the limits depend on the values of the other parameters. Besides this linear constraint, at the same time the non-linear inequality $k_{1}^{2} \geq 4 k_{0} k_{2}$ must also be taken into account. If the standard functions are used for the problem at hand, then the algorithm will typically step outside the feasible region leading to numerical errors. The function solnp from the R-package Rsolnp (Ghalanos and Theussl, 2014) solves general nonlinear programming problems and allows for inequalities and non-linear equalities, and can handle our maximization problem. Figure 1 also represents the standard relationships whose compositions are given in Table 2. All these relationships are at the edge of the feasible region.

The parabola delimiting the feasible region described by $k_{1}^{2}=4 k_{0} k_{2}$ can be recognized as coinciding with the second dimension of the isometric log-ratio coordinates (Egozcue et al., 2003) of composition $\mathbf{k}=\left(k_{0}, k_{1}, k_{2}\right)$. The same parabola is also of relevance in studies of Hardy-Weinberg equilibrium (Graffelman \& Egozcue, 2011), but with a different genetic interpretation of the composition. This suggests that we might re-parametrize the likelihood in terms of the isometric log-ratio coordinates in order to obtain a rectangular domain for the likelihood. This simplifies the maximization problem, as it can now be solved using R's general-purpose optimization routines optim and nlminb.

We use the isometric log-ratio transformation, calculating the coordinates as follows:

$$
\begin{equation*}
z_{1}=\frac{1}{\sqrt{2}} \ln \left(\frac{k_{0}}{k_{2}}\right), \quad z_{2}=\frac{1}{\sqrt{6}} \ln \left(\frac{k_{0} k_{2}}{k_{1}^{2}}\right) . \tag{4}
\end{equation*}
$$

The inverse relationships are given by:

$$
\begin{equation*}
\left(k_{0}, k_{1}, k_{2}\right)=\mathcal{C}\left(e^{\sqrt{2} z_{1}}, e^{\frac{1}{2} \sqrt{2} z_{1}-\frac{1}{2} \sqrt{6} z_{2}}, 1\right) \tag{5}
\end{equation*}
$$

where $\mathcal{C}$ is the closure operator. This gives the re-parameterized log-likelihood:

$$
\begin{equation*}
l\left(z_{1}, z_{2} \mid G_{1} \cap G_{2}\right)=\sum_{i=1}^{n}\left(\ln \left(d_{0 i} e^{\sqrt{2} z_{1}}+d_{1 i} e^{\frac{1}{2} \sqrt{2} z_{1}-\frac{1}{2} \sqrt{6} z_{2}}+d_{2 i}\right)-\ln \left(1+e^{\sqrt{2} z_{1}}+e^{\frac{1}{2} \sqrt{2} z_{1}-\frac{1}{2} \sqrt{6} z_{2}}\right)\right) . \tag{6}
\end{equation*}
$$

The non-linear constraint $k_{1}^{2} \geq 4 k_{0} k_{2}$ becomes a linear inequality for the second ilr-coordinate:

$$
\begin{equation*}
z_{2} \leq-\sqrt{\frac{2}{3}} \ln (2) \tag{7}
\end{equation*}
$$

The domain of the re-parameterized log-likelihood is shown in Figure 1B.


Figure 1: Domain of the likelihood function (grey) in a ternary plot representation (A) and in isometric log-ratio coordinates (B).

## 4 Examples

We use data from the Mexican population of phase III of the HapMap project (The International HapMap Consortium, 2010) to illustrate the estimation of IBD probabilities in log-ratio coordinates. The data consist of genotype information of 86 individuals (mostly parent-offspring trios) from Los Angeles of Mexican ancestry. Several scholars have analyzed these data, and many undocumented family relationships have been reported (Pemberton et al., 2010; Stevens et al., 2012). We filtered single nucleotide polymorphisms (SNPs) from the genome-wide HapMap database as follows. SNPs significant in a chi-square test for Hardy-Weinberg equilibrium ( $\alpha=0.05$ ) were excluded to avoid possible genotyping error. SNPs with a minor allele frequency below 0.4 were also excluded in order to guarantee a set of sufficiently polymorphic markers. We sampled 5.000 SNPs at random from this subset, and consider the resulting dataset as a set of approximately independent and highly polymorphic markers. We consider two pairs chosen from this database as an example.

The first pair is a presumably unrelated pair of individuals with identifiers NA19662 and NA19685. This pair was inferred to be a FS pair (Pemberton et al., 2010) using the program RELPAIR (Epstein et al., 2000). Here we re-analyze the relationship of this pair using ML estimation of the IBD probabilities, with initial point $(0.575,0.400,0.025)$. The iteration history and log-likelihood ( $l$ ) are given in Table 4 for the maximization of Equation (3) with solnp and for the maximization in ilr-coordinates (Eq. (6)) with nlminb. Both algorithms converge to the same maximum. Back-transformation of the final ilr-coordinates ( $-0.1143,-0.6601$ ) gives the same estimates $\hat{k}_{0}=0.217, \hat{k}_{1}=0.528$ and $\hat{k}_{2}=0.255$, which confirms the hypothesis of a FS pair. The level

Table 4: ML estimation of IBD probabilities of a FS pair, using 5.000 SNPs, with initial point ( $0.575,0.400,0.025$ ). Iteration histories with the log-likelihood $(l)$ for maximization in original and in log-ratio coordinates by solnp and nlminb respectively.

| solnp | It. | $l$ | $\hat{k}_{0}$ | $\hat{k}_{1}$ | $\hat{k}_{2}$ |
| ---: | ---: | ---: | ---: | ---: | ---: |
|  | 1 | -9483.1290 | 0.41422 | 0.48104 | 0.10474 |
|  | 2 | -9368.1777 | 0.18452 | 0.56753 | 0.24796 |
|  | 3 | -9366.4621 | 0.21746 | 0.52776 | 0.25478 |
|  | 4 | -9366.4615 | 0.21697 | 0.52798 | 0.25505 |
|  | 5 | -9366.4615 | 0.21697 | 0.52798 | 0.25505 |
| nlminb | It. | $l$ | $\hat{z}_{1}$ | $\hat{z}_{2}$ |  |
|  | 0 | -9671.5480 | 2.217130 | -0.983749 |  |
|  | 1 | -9503.2604 | 1.528570 | -1.708930 |  |
| 2 | -9446.4083 | 0.538662 | -1.567190 |  |  |
| 3 | -9415.8636 | 0.214144 | -1.361630 |  |  |
| 4 | -9367.0124 | -0.184897 | -0.705093 |  |  |
|  | -9366.5961 | -0.108992 | -0.696269 |  |  |
| 6 | -9366.4802 | -0.127274 | -0.667618 |  |  |
| 7 | -9366.4659 | -0.113317 | -0.666588 |  |  |
| 8 | -9366.4622 | -0.116864 | -0.661281 |  |  |
| 9 | -9366.4617 | -0.114033 | -0.661312 |  |  |
| 10 | -9366.4615 | -0.114846 | -0.660250 |  |  |
| 11 | -9366.4615 | -0.114261 | -0.660282 |  |  |
| 12 | -9366.4615 | -0.114416 | -0.660107 |  |  |
| 13 | -9366.4615 | -0.114323 | -0.660105 |  |  |
| 14 | -9366.4615 | -0.114323 | -0.660105 |  |  |

curves of the log-likelihood function are shown in Figure 2A and show the maximum as an interior point.

We consider a second example of another undocumented relationship between NA19660 and NA19685 that was inferred to be a parent-offspring (PO) pair (Pemberton et al., 2010). This example differs from the previous one because now the reference relationship ( PO with $k_{0}=0, k_{1}=1$ and $k_{2}=0$ ) is outside the simplex. Results for the maximization in original and log-ratio coordinates are shown in Table 5 below.

Table 5: ML estimation of IBD probabilities of a PO pair, using 5.000 SNPs, with initial point ( $0.575,0.400,0.025$ ). Iteration histories for maximization of the likelihood function ( $l$ ) in original and in log-ratio coordinates by solnp and nlminb respectively.

| solnp | It. | $l$ | $\hat{k}_{0}$ | $\hat{k}_{1}$ | $\hat{k}_{2}$ |
| ---: | ---: | ---: | ---: | ---: | ---: |
|  | 1 | -9489.1607 | 0.29473 | 0.70527 | 0.00000 |
|  | 2 | -9320.4669 | 0.00184 | 0.99816 | 0.00000 |
|  | 3 | -9320.4669 | 0.00184 | 0.99816 | 0.00000 |
| nlminb | It. | $l$ | $\hat{z}_{1}$ | $\hat{z}_{2}$ |  |
|  | 0 | -9680.5924 | 2.21713 | -0.98375 |  |
|  | 1 | -9324.1179 | 0.24058 | -3.86971 |  |
|  | 2 | -9322.1662 | 1.56428 | -4.90716 |  |
|  | 3 | -9321.5402 | 2.13094 | -5.41660 |  |
|  | 4 | -9320.7389 | 3.37902 | -6.55765 |  |
|  | 5 | -9320.5206 | 4.20767 | -7.31837 |  |
|  | 6 | -9320.4712 | 4.73722 | -7.80468 |  |
|  | 7 | -9320.4672 | 4.91901 | -7.97160 |  |
|  | 8 | -9320.4671 | 4.94894 | -7.99905 |  |
|  | -9320.4671 | 4.95042 | -8.00037 |  |  |

Function solnp gives estimates $\hat{k}_{0}=0.0018, \hat{k}_{1}=0.9982$ and $\hat{k}_{2}=0.0000$ which suggests a PO pair. The maximization in log-ratio coordinates shows that $z_{1}$ increases and $z_{2}$ decreases until the change in the log-likelihood drops below the tolerance used. Back-transformation of the coordinates gives the result $\hat{k}_{0}=0.0018, \hat{k}_{1}=0.9982$ and $\hat{k}_{2}=0.0000$ which coincides with the estimates obtained by solnp. Figure 2B shows the level curves of the log-likelihood function together with the maximum found (marked with a red dot). As $z_{2}$ decreases at some point the log-likelihood function becomes very flat.


Figure 2: Level curves of the log-likelihood function in ilr-coordinates for a FS pair and a PO pair.
For the FS pair, different initial points were used and all converged to the same maximum. For
the PO pair, different starting points often give different solutions in ilr-coordinates. The maxima found according to the iteration histories in Tables 4 and 5 are marked by a red point. For the PO pair, the solutions obtained from three different additional initial points are shown by black triangles. In ilr-coordinates these solutions differ, but all correspond to an area where the loglikelihood function is very flat. Back-transformation of all solutions gives however, up to five decimals, the same IBD probabilities.

## 5 Discussion

In this contribution we have shown that isometric log-ratio coordinates can be used to simplify a genetic maximization problem that has the simplex or a subset of the simplex as its domain. In statistics there are more likelihood functions that are subject to a unit sum constraint on the parameters, and that may possibly be simpler to maximize in log-ratio coordinates, such as likelihoods based on the multinomial distribution. Maximization of likelihoods in coordinates may therefore have a wider applicability then suggested by the specific genetic problem dealt with here. The main advantage is that the irregular domains of the likelihood function in the simplex can become rectangular when expressed in ilr-coordinates. Most standard maximization functions can then handle the maximization problem, whereas specialized software is needed for maximizing the likelihood function over a subset of the simplex.

We note that the theoretical IBD probabilities for the most common family relationships fall on the edge of the simplex (See Figure 1A). Only the FS relationship is an interior point of the simplex and all other relationships from Table 2 are outside the simplex because zeros are not admitted. Thus, when maximizing in coordinates, these theoretical probabilities can never be attained. In practice the log-ratio coordinates tend to extreme values for these relationships, and convergence can be slow because the log-likelihood function flattens. Setting an adequate tolerance criterion may help to speed up the convergence.

The ilr-coordinates of the numerical solution found for relationships on the edge of the simplex can vary considerably (see Figure 2B). However, when back-transformed to IBD probabilities the corresponding relationships can be inferred. In this respect, we note that large negative values for $z_{1}$ combined with $z_{2}$ at its maximum of $-\sqrt{2 / 3} \ln (2)$ point to a MZ pair, and large positive values for $z_{1}$ combined with $z_{2}$ at its maximum of $-\sqrt{2 / 3} \ln (2)$ suggest an unrelated pair (UN). Avuncular pairs (AV, aunt-niece, etc.), grand-parent-grandchild pairs (GG) and half-sibs (HS) can not be distinguished because they all have the same Cotterman coefficients. With the isometric log-ratio transformation proposed here, these relationships are characterized by a positive $z_{1}$ and proportionality between the log-ratio coordinates $\left(z_{1}=-\sqrt{3} z_{2}\right)$.

## Acknowledgments

This study was supported by grant CODARSS MTM2012-33236 of the Spanish Ministry of Education and Science.

## References

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

Epstein, M., Duren, W. and Boehnke, M. (2000). Improved inference of relationship for pairs of individuals. American Journal of Human Genetics 67, pp. 1219-1231.

Ghalanos, A. and Theussl, S. (2014). Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method. R package version 1.15.

Graffelman, J. and Egozcue, J.J. (2011). Hardy-Weinberg equilibrium: a nonparametric compositional approach. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), Compositional Data Analysis: Theory and Applications, pp. 208-217. John Wiley \& Sons, Ltd.

Pemberton, T.J., Wang, C., Li, J.Z. and Rosenberg, N.A. (2010), Inference of unexpected genetic relatedness among individuals in HapMap Phase III. American Journal of Human Genetics 87, pp. 457-464.

R Core Team (2014). R: A Language and Environment for Statistical Computing. http://www.Rproject.org/ R Foundation for Statistical Computing, Vienna, Austria.

Stevens, E.L., Baugher, J.D., Shirley, M.D., Frelin, L.P. and Pevsner, J. (2012). Unexpected Relationships and Inbreeding in HapMap Phase III Populations. PLoS ONE 7(11), e49575. doi:10.1371/journal.pone. 0049575
Thompson, E.A. (1991). Estimation of relationships from genetic data. In C.R. Rao and R. Chakraborty (Eds.), Handbook of Statistics. Elsevier Science: Amsterdam, Vol 8, pp. 255-269.

The International HapMap Consortium (2010). Integrating common and rare genetic variation in diverse human populations. Nature 467 , pp. 52-58.

Weir, B.S., Anderson, A.D. and Hepler, A.B. (2006). Genetic relatedness analysis: modern data and new challenges Nature Reviews Genetics 7(10), pp. 771-780.

# Biplots for Compositional Regression 

J. Graffelman ${ }^{1}$ and R. Tolosana-Delgado ${ }^{2}$<br>${ }^{1}$ Department of Statistics and Operations Research, Universitat Politècnica de Catalunya, Barcelona, Spain; jan.graffelman@upc.edu<br>${ }^{2}$ Department of Modeling and Evaluation, Helmholtz-Institut Freiberg für Ressourcentechnologie, Freiberg, Germany; r.tolosana@hzdr.de


#### Abstract

In the statistical modeling of compositions, the latter can appear as response or explanatory variables, or even in both roles simultaneously. Such modeling is of multivariate nature because compositions consists of vectors of $D$ elements, with potentially large $D$. Prior to fitting a model, the compositions are typically first transformed by a log-ratio transformation. The transformed compositions remain, however, of multivariate nature. Several techniques from the field of multivariate analysis can be used to analyze the data of which we mention multivariate regression, canonical correlation analysis and redundancy analysis. In multivariate regression the interpretation of the results is complicated by the large number of regression coefficients obtained, and collected in the matrix of regression coefficients $\mathbf{B}_{p \times q}$ with $p$ explanatory and $q$ response variables. Van den Boogaart and Tolosana-Delgado (2013) applied the singular value decomposition of the matrix of regression coefficients in order to find rank two approximations to $\mathbf{B}$ so that biplots can be constructed that summarize the relationships between the variables. However, the regression coefficients depend on the scale of the explanatory variables and their variances can be of very different order of magnitude. It is thus more natural to "standardize" the regression coefficient with a Mahanalobis-like transformation prior to biplot construction. Such an approach amounts to a redundancy analysis (also known as reduced-rank regression) of the data, and biplots for this method are discussed in detail by Ter Braak \& Looman (1994). In this contribution we develop and apply redundancy analysis in a compositional context, discussing its biplots with empirical geological data and its compositional geometric interpretation.


Key words: Redundancy analysis; Multivariate regression; Biplot; Conditional biplot; centred log-ratio (clr) transformation.

## 1 Introduction

Compositional data analysis is inherently of multivariate nature, because most compositions consist of more than 2 parts. In order to summarize the relationships between the components, principal component analysis of the log-ratios of the compositions is often employed. In principal component analysis, all variables have the same status and no distinction is made between explanatory and response variables. Biplots that summarize the compositional data matrix have been described by Aitchison and Greenacre (2002). However, in many studies compositions are involved in a modeling framework, where compositions can take the role of explanatory variables, of response variables, or even of both. Because compositions are multivariate data sets, such modeling requires multivariate techniques that can summarize the relationships between two sets of variables such as multivariate regression, canonical correlation analysis or redundancy analysis. Biplots have great potential in this context, because they can summarize the relationships between the variables by means of a biplot of the between-set correlation matrix or of the matrix of regression coefficients. Biplots for redundancy analysis (Ter Braak and Looman, 1994) and canonical correlation analysis (Graffelman, 2005) are well described in the literature, but up to date, little effort has been made to adapt such biplots to the particularities of linear models for compositional data. The purpose of this contribution is to explore the use of biplots for visualizing the relationships between a compositional response variable and a set of non-compositional explanatory variables. In the remainder of this contribution we develop some theory for constructing biplots of regression coefficients (Section 2) and give an example with a geochemical data set (Section 3).

## 2 Theory

The relationships between the compositions $(\mathbf{Y})$ and the explanatory variables are summarized by the matrix of regression coefficients $\mathbf{B}$. The multivariate multiple regression is given by

$$
\mathbf{Y}=\mathbf{X} \boldsymbol{\beta}+\boldsymbol{\varepsilon}
$$

where $\mathbf{Y}(n \times q)$ contains the response variables, $\mathbf{X}(n \times p)$ is the matrix with explanatory variables, $\boldsymbol{\beta}$ the $p \times q$ matrix of regression coefficients and $\boldsymbol{\varepsilon}$ the $n \times q$ matrix of errors. The ordinary least squares estimator for $\boldsymbol{\beta}$ is given by:

$$
\begin{equation*}
\mathbf{B}=\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-1} \mathbf{X}^{\prime} \mathbf{Y} \tag{1}
\end{equation*}
$$

In this multivariate regression we assume that $\mathbf{X}$ contains no leading column of ones, because we are not interested in representing the intercept term. The $X$-variables are assumed to be columncentred or eventually standardized if there are differences in scale of the $X$-variables, and we take the centred log-ratio coordinates (clr) of the compositions as response variables. We assume the clr transformed response variables in matrix $\mathbf{Y}$ to be column-centred as well. Working with centered variables effectively removes the intercept term from the model. A biplot of the matrix of regression coefficients can be constructed by using the singular value decomposition of $\mathbf{B}$ :

$$
\mathbf{B}=\mathbf{U D V}^{\prime}
$$

with $\mathbf{U}^{\prime} \mathbf{U}=\mathbf{I}$ and $\mathbf{V}^{\prime} \mathbf{V}=\mathbf{I}$, and where $\mathbf{D}$ contains the singular values in non-increasing order of magnitude. The full space of the solution will, in general, contain $k=\min (p, q)$ dimensions. However, because the clr coordinates are subject to a zero sum constraint, the regression coefficients in Equation (1) also have row sum zero and this reduces the rank of the solution to $k=\min (p, q-1)$. Principal and standard coordinates for $X$-variables are given by

$$
\mathbf{F}_{p}=\mathbf{U D}, \quad \mathbf{F}_{s}=\mathbf{U}
$$

and principal and standard coordinates for the $Y$-variables in the biplot are given by

$$
\mathbf{G}_{p}=\mathbf{V D}, \quad \mathbf{G}_{s}=\mathbf{V}
$$

A biplot of the regression coefficients is obtained by jointly plotting the first two columns of $\mathbf{F}_{p}$ with the first two columns of $\mathbf{G}_{s}$, or conversely the first two columns of $\mathbf{F}_{s}$ with the first two columns of $\mathbf{G}_{p}$. These are the most common biplot scalings, though other choices are possible, depending on how one distributes the singular values over the row and column markers. The goodness-of-fit of a two-dimensional approximation to $\mathbf{B}$ is given by the ratio of the sum of the first two squared singular values (also eigenvalues of $\mathbf{B}^{\prime} \mathbf{B}$ ) with respect to the total sum of squares.

$$
\left(d_{1}^{2}+d_{2}^{2}\right) / \sum_{i=1}^{k} d_{i}^{2}
$$

and corresponds to the total sum-of-squares of the entries in $\mathbf{B}$ explained by a two-dimensional approximation. A shortcoming of this biplot is that it only shows the relationships between $X$ and $Y$-variables, but that the samples of the data matrix are not represented. We propose to enrich the biplot with a representation of the samples. The same problem of mapping samples into a biplot has been described before for canonical correlation analysis, where biplots of the between-set correlation matrix can be amplified with points that represent the original samples (Graffelman, 2005). The representation of the samples in a second step can be done in several ways, depending on whether we want to display the original compositions, the original $X$-variables or both. If we focus on explaining the compositions, it may be most natural to add the original compositions to the biplot. This means that we construct a conditional biplot of $\mathbf{Y}=\mathbf{H}_{y} \mathbf{G}_{s}^{\prime}$, where the column markers of that biplot are constrained to be equal to $\mathbf{G}_{s}$ ( or $\mathbf{G}_{p}$ ). The coordinates for the compositions are calculated by

$$
\begin{equation*}
\mathbf{H}_{y}=\mathbf{Y} \mathbf{G}_{s}\left(\mathbf{G}_{s}^{\prime} \mathbf{G}_{s}\right)^{-1} \tag{2}
\end{equation*}
$$

The compositions are represented in the biplot by plotting the first two columns of matrix $\mathbf{H}_{y}$. Alternatively, the samples of the $X$-measurements may be added to the biplot by constructing a conditional biplot of $\mathbf{X}=\mathbf{H}_{x} \mathbf{F}_{p}^{\prime}$, where $\mathbf{H}_{x}$ is obtained by

$$
\begin{equation*}
\mathbf{H}_{x}=\mathbf{X} \mathbf{F}_{p}\left(\mathbf{F}_{p}^{\prime} \mathbf{F}_{p}\right)^{-1} \tag{3}
\end{equation*}
$$

In both cases we use regression results to fit supplementary information to a given biplot (Graffelman \& Aluja-Banet, 2003). Predictors in multiple and multivariate linear regression are often correlated. The correlation structure of the predictors is ignored in the biplot discussed so far. Redundancy analysis (RDA), also known as reduced-rank regression, offers an alternative way for obtaining a graphical representation of the matrix of regression coefficients that takes correlated predictors into account. RDA is due to Van den Wollenberg (1977), and biplots for RDA have been proposed and described in detail by Ter Braak and Looman (1994). RDA has become a popular method in ecology. A well-known result in multiple regression is that the covariance matrix of the regression coefficients is given by $\sigma^{2}\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-1}$. Transformation of the matrix of regression coefficients by premultiplying by $\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{\frac{1}{2}}$ can be used to account for correlations among the $X$-variables, and this is precisely what is done in RDA. With this transformation, we perform the singular value decomposition of the transformed regressions coefficients given by

$$
\mathbf{B}=\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{\frac{1}{2}}\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-1} \mathbf{X}^{\prime} \mathbf{Y}=\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-\frac{1}{2}} \mathbf{X}^{\prime} \mathbf{Y}=\mathbf{U D V}^{\prime}
$$

and biplot coordinates for explanatory $X$-variables are obtained by

$$
\mathbf{F}_{p}=\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-\frac{1}{2}} \mathbf{U D}, \quad \mathbf{F}_{s}=\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-\frac{1}{2}} \mathbf{U}
$$

while biplot coordinates for the response compositions $Y$ are given by

$$
\mathbf{G}_{p}=\mathbf{V D}, \quad \mathbf{G}_{s}=\mathbf{V}
$$

This gives a generalized least squares (GLS) fit to the matrix of regression coefficients, as

$$
\mathbf{F}_{p} \mathbf{G}_{s}^{\prime}=\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-\frac{1}{2}} \mathbf{U D V}^{\prime}=\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-1} \mathbf{X}^{\prime} \mathbf{Y}=\mathbf{B}
$$

Note that, in the scaling with standard row coordinates, the full-space solution recovers the covariance matrix between the regression coefficients because

$$
\mathbf{F}_{s} \mathbf{F}_{s}^{\prime}=\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-\frac{1}{2}} \mathbf{U}^{\prime}\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-\frac{1}{2}}=\left(\mathbf{X}^{\prime} \mathbf{X}\right)^{-1}
$$

This property can be used to detect multicollinearity.

## 3 Example

We use a dataset of proportions of four petrographic types of sediment grains as an example. This dataset was collected by Grantham and Velbel (1988), and its analysis is discussed by Van den Boogaart and Tolosana-Delgado (Chapter 5, 2013). The data concern 72 samples of 4 part compositions consisting of polymineralic rock $\left(R_{f}\right)$, grains of a single quartz crystal $\left(Q_{m}\right)$, grains containing several quartz crystals $\left(Q_{p}\right)$ and grains of mica $(M)$. The explanatory variables for these compositions concern the grain size (fine, medium or coarse), position (north or south), discharge and relief. The samples come from low-order streams of the Coweeta Basin in North Carolina (USA). More geological details of the study are given by Grantham and Velbel (1988) and Tolosana-Delgado and van den Boogaart (Chapter 26, 2011). A biplot of the regression coefficients obtained by OLS is shown in Figure 1.


Figure 1: Biplot of the regression coefficients $\mathbf{B}$ for a geochemical data set.

The biplot suggests the ratio $Q_{p}: M$ to increase as the grain size increases from fine to medium and coarse. The ratio $Q_{m}: M$ seems to increase from north to south, or alternatively decrease with increasing discharge. Orthogonality allows to assess as well lack of dependence between logratios and explanatory variables. In this sense, the biplot suggests that the ratio $Q_{m}: M$ might be not depend on grain size, or that discharge does not change the ratio $Q m: Q p$. Goodness-of-fit statistics are shown in Table 1.

Table 1 shows that a two-dimensional biplot accounts for $98 \%$ of the total sum of squares, which is an excellent fit. Note that one zero eigenvalue appears as explained in Section 2. We proceed to enhance the biplot in Figure 1 with the original compositions, by using the regression formula

| Table 1: Goodness-of-fit statistics for B |  |  |  |  |
| :--- | :---: | :---: | :---: | :---: |
|  | 1 | 2 | 3 | 4 |
| eigenvalue | 2.30 | 0.20 | 0.06 | 0.00 |
| fraction | 0.90 | 0.08 | 0.02 | 0.00 |
| cumulative | 0.90 | 0.98 | 1.00 | 1.00 |

in Equation (2). This produces the biplot in the left panel of Figure 2, where the compositions appear coloured according to the grain size. Interpretation of the compositions should be made with respect to the log-ratios only. This biplot reveals that there are 3 southern samples which are outliers with respect to the $Q_{m}: M$ ratio. In the right panel of Figure 2 we added the samples of the $X$-measurements using Equation (3). Again we coloured samples according to the grain size, and here used different symbols for northern and southern samples. The plot in the right panels appears to contain fewer samples, but this is actually not the case. The $X$-variables contain many repeated values, and for this reason many samples map to the same point in the biplot. This biplot confirms the horizontal axis to be a grain size gradient, and the vertical axis to be a north-south difference plus a discharge gradient.


Figure 2: Biplot of the regression coefficients enriched with the original compositions (left) or enriched with the original $X$-measurements (right).

Finally, we present the biplot obtained by redundancy analysis in Figure 3. This plot has a goodness-of-fit of 0.97 and closely resembles the previous biplots. This is probably due to the fact that the $X$-measurements in this study have relatively little correlation. There are nevertheless some details in the RDA that deserve attention. The South and Coarse vector are almost orthogonal. This means the corresponding regression coefficients are virtually uncorrelated, and that the effect of these two variables on the compositions can be interpreted separately. Note that this fact is in this case enhanced by the design of the data, which was constructed perfectly balanced between the two-level factor North-South and the three-level factor Fine-Medium-Coarse. The most notable change in comparison with the previous analysis is the orientation of the variable vector Relief. The regression coefficients of Relief, Discharge and the North-South difference are correlated to some extent, and this suggests that their effects on the compositions are confounded to some degree. RDA is specially suitable to detect such features, as it also displays the correlation matrix of the regression coefficients.


Figure 3: Biplot of the regression coefficients obtained by redundancy analysis.

## 4 Conclusion

The main result of this contribution is that we have managed to summarize the relationships between compositions and set of explanatory variables by means of a biplot of the regression coefficients, enhancing the latter with a representation of the original compositions or $X$-variables of the data matrix. We think redundancy analysis and canonical correlation analysis are promising multivariate tools for compositional data analysis that deserve to be explored and studied in more detail in the context of the log-ratio approach. RDA biplots have the potential of detecting multicollinearity problems. These tools are particularly useful to study potential subcompositional independence, i.e. to detect subcompositions that do not (linearly) depend on certain explanatory covariables.

## Acknowledgments

This study was supported by grant CODARSS MTM2012-33236 of the Spanish Ministry of Education and Science.

## References

Aitchison, J. \& Greenacre, M. (2002). Biplots for compositional data. Journal of the Royal Statistical Society, Series C (Applied Statistics) 51(4), pp. 375-392.

Graffelman, J. (2005). Enriched biplots for canonical correlation analysis. Journal of Applied Statistics 32(2), pp. 173-188.

Graffelman, J. \& Aluja-Banet, T. (2003). Optimal representation of supplementary variables in biplots from principal component analysis and correspondence analysis. Biometrical Journal $45(4)$, pp. 491-509.

Grantham, J.H. \& Velbel, M.A. (1988). The influence of climate and topography on rock-fragment abundance in modern fluvial sands of the southern blue ridge mountains, North Carolina. Journal of Sedimentary Petrology 58, pp. 219-227. Springer.

Ter Braak, C.J.F \& Looman, C.W.N. (1994). Biplots in reduced-rank regression. Biometrical Journal 36(8), pp. 983-1003.

Tolosana-Delgado, R. \& van den Boogaart, K.G. (2011). Linear models with Compositions in R. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), Compositional Data Analysis: Theory and Applications. Chapter 26, pp. 356-371. John Wiley \& Sons.

Van den Boogaart, K.G. \& Tolosana-Delgado, R. (2013). Analyzing Compositional Data with R. Springer.

Van den Wollenberg, A.L. (1977) Redundancy Analysis, an alternative for canonical correlation analysis. Psychometrika 42(2), pp. 207-219.

Recognizing and Validating Structural Processes in Geochemical Data E.C. Grunsky, B.A. Kjarsgaard ${ }^{1}$<br>${ }^{1}$ Geological Survey of Canada, Ottawa, Canada K1A 0E8


#### Abstract

Geochemical data are compositional in nature and are subject to the problems typically associated with data that is restricted in real positive numbers space, the simplex. Geochemistry is a proxy for mineralogy, and minerals are comprised of atomically ordered structures that define the placement and abundance of elements with a mineral lattice structure. The arrangement of elements within one or more minerals that comprise rocks, soils and aqueous solutions define a linear model in terms of their geochemical expression. When methods such as principal component analysis are applied to multi-element geochemical data, the dominant components generally reflects features related to mineralogy and describe geologic processes that are both independent and partially codependent. The dominant principal components can be used as a filter to eliminate noise or under-sampled processes in the data. These dominant components can be used to create predictive geological maps, or maps displaying recognizable geochemical processes.


Kew words: geochemistry, classification, lithologic prediction, compositional data analysis

## 1 Introduction

The compositional nature of geochemical data requires care when considering relationships between the elemental, oxide or molecular constituents that define a composition. The use of ratios are essential when making comparisons between elements in systems such as igneous fractionation (Pearce, 1969) and the use of logratios is essential when measuring moments such as variance/covariance in examination of data derived from geochemical surveys (Aitchison, 1986; Egozcue et al, 2003; Buccianti et al., 2006; Pawlowsky and Buccianti, 2011).
The relationships between the elements of geochemical data are governed by "natural laws" (Aitchison, 1999) and specifically by stoichiometry and thereby imposing structure within the data. Geochemical data are not the only data that exhibit structure. Grunsky and Bacon-Shone (2011) have shown that geochemical patterns and trends are closely related to the stoichiometric constraints of minerals.
To effectively interpret geochemical data, a two-phase approach is suggested; that of process discovery, followed by process validation. This tactic identifies geochemical/geological processes that exist in the data but are not obvious unless robust statistical methods are utilized. The process discovery phase is most effective when carried out using a multivariate approach. Linear combinations of elements related by stoichiometry are generally expressed as strong patterns, whilst random patterns and under-sampled processes show weak or uninterpretable patterns. This was initially demonstrated by Grunsky et al. (2012a) using multi-element lake sediment geochemical data from the Melville Peninsula, Nunavut, Canada.
Two examples are presented. The first demonstrates the usefulness of compositional data analysis combined with multivariate statistical analysis for process discovery and validation from drill core geochemistry from the Star kimberlite in Saskatchewan, Canada (Grunsky \& Kjarsgaard, 2008). Distinct geochemical kimberlite phases can be statistically identified using this approach and lead to efficiencies in the economic evaluation of kimberlite for diamonds.
The second example is the evaluation of lake sediment geochemistry obtained from a regional geochemical survey (McCurdy et al, 2013). The results presented here are from a campaign to re-analyze sample pulps using modern analytical methods including Inductively Coupled Plasma Mass Spectrometry (ICP-MS). In cases where elements have been analyzed using two or more methods, the elements were evaluated in terms of detection limit suitability and visual examination of the correlation of the element with each method. The elements B, Ge, In, Pd, Re, Ta and Te were dropped due to large numbers of observations that were reported at less than the lower detection limit (lld).
One of the primary purposes of geochemical data analysis is the recognition of geochemical/geological processes. Processes are recognized by a continuum of variable responses and the relative increase/decrease of these variables. The presence of censored data (values < lld) can, in some cases, affect the results of a process recognition investigation. In the compositional data analysis framework, early on Aitchison (1986) recognized the problem of censored data. Martin-Fernandez et al. (2003) discusses various replacement options based on the nature of the censored data. Recognizing the difference between missing (i.e. no data) values and censored (< lld) data is crucial in deciding how a replacement value, if any, should be estimated. Details on the methods used are described below.

## 2 Methods

### 2.1 Process Discovery

Process discovery involves the use of unsupervised multivariate methods such as principal component analysis (PCA), multi-dimensional scaling and Random Forests, to name a few. Process discovery with geochemical data reflects the recognition of linear models that reflect the stoichiometry of rock forming minerals and subsequent processes that modify mineral structures (hydrothermal fluids, weathering, groundwater). Additional processes such as gravitational effects such as sorting arrange minerals according to the energy of the environment and mineral density. The chemistry of minerals are governed by stoichiometry and the relationships are easily described within the simplex. Geoscientists have long recognized that many geochemical processes can be clearly described using cation ratios that reflect the stoichiometric balances of minerals during formation (e.g. Pearce, 1968).
Many geochemical datasets contain values that are reported at less than the lower limit of detection and these values are generally termed "censored". The estimation of statistical parameters can be severely affected by censored data and it is useful to find a replacement value that does not bias the estimate of the statistical moments. The R package "zCompositions" with the function (lrEM) (Palarea-Albaladejo et al., 2008) was used to determine suitable replacement values for several of the elements. Missing values can
be the result of transcription omissions, or result from studies that use multiple analytical methods in which the suite of analyzed elements are not identical, require different treatment (Grunsky, 2010) and is not discussed further in this study.
When mixtures of minerals (rocks, soils, glacial till, stream sediments) are analyzed for their geochemistry there may be multiple linear processes embedded in the resulting geochemical analyses. Techniques such as principal component analysis can be effectively used to identify these processes. A method of PCA used in this study is a combined R-mode/Q-mode PCA as documented in Grunsky (2001).

An essential part of the process discovery phase is a suitable choice of coordinates to overcome the problem of closure. The logcentred transform (Aitchison, 1986) is a suitable transform for the evaluation of geochemical data. The resulting principal components are orthonormal and reflect linear processes related to stoichiometry. The components are ideal for subsequent process validation. The methods used in this study were applied in the R programming environment (R Core Team, 2014) and are documented below. Enhancements to the visualization of groups of data on scatterplots were made using the R library "cluster" and the function "ellipsoidhull", which creates a convex hull around specified groups of data.

### 2.2 Process Validation

Process validation is the methodology employed to verify that a geochemical composition (response) is associated with identified processes. The processes can be in the form of e.g., lithology, soil character, ecosystem properties, climate, or deeply buried tectonic assemblages. Validation can be in the form of an estimate of likelihood that a composition can be assigned membership to one of the identified processes. This is typically done through the assignment of class identifier or a measure of probability. Assignment of class membership can be done through the application of techniques such discriminant analysis, logistic regression, neural networks or, Random Forests. There are many other methods available.
An essential part of process validation is the selection of variables that enable efficient classification. Efficient classification involves the selection of variables that maximize the differences between the different classes and minimizes the amount of overlap due to noise or unresolved processes in the data. Within the context of compositional data, variables that are selected for classification require transformation to logratio coordinates. The additive logratio (alr) or the isometric logratio (ilr) are equally effective for the implementation of classification procedures. The logcentred (clr) transform is not suitable because the covariance matrix of these coordinates is singular. However, analysis of variance applied to logcentred data enables recognition of the compositional variables (elements) that are most effective for distinguishing between the classes. Choosing an effective alr transform (choice of suitable divisor) or balances for the ilr transform is not a trivial task. However, analysis of variance applied to the principal components derived from the logcentred transform can be highly effective at discriminating between the different classes. Subsequent classification can be carried out with far fewer variables based on principal components derived from the logcentred transform.
Initially, an analysis of variance ( R function "aov") is applied to the logcentred elements or the principal components derived from the logcentred data. In this study, as explained below, the principal components were used to in the analysis of variance.
Classification results can be conveniently expressed as direct class assignment or posterior probabilities in the form of forced class allocation, or as class typicality. Forced class allocation assigns a posterior probability based on the shortest Mahalanobis distance of a compositional observation from the compositional centroid of each class. Class typicality measures the Mahalanobis distance from each class and assigns a posterior probability based on the F-distribution. This approach can result with an observation having a zero posterior probability for all classes, indicating that the composition is not close to the compositions defined by the class compositional centroids. The results shown for this study used the R function, "lda", from which the posterior probabilities were estimated.
Classification accuracies can be assessed through the generation of tables that show the accuracy and errors measured from the estimated classes against the initial classes in the training sets used for the classification.

### 2.2 Geospatial Coherence

The results from the classification of materials gathered from a geochemical survey should bear a geospatial resemblance to the area sampled. The creation of maps are part of the process validation procedure. If a geospatial rendering of a posterior probability shows no spatial coherence, then it is unlikely that the classification is meaningful. The most effective way to test this is through the generation
and modelling of semi-variograms that describe the spatial continuity of a specific class based on the posterior probabilities. If meaningful semi-variograms can be created, then geospatial maps of the posterior probabilities can be generated through interpolation using the kriging process. Maps of posterior probabilities may show low overall values but still be spatially coherent. This is also reflected in the classification accuracy matrix that indicates the extent of classification overlap between classes.
Geospatial analysis methodology described by Bivand et al. (2013) and the gstat package (Pebesma, 2004) were used to generate the geostatistical parameters and images of the principal components and posterior probabilities from kriging.

## 3 Examples

### 3.1 Process Discovery and Validation of Diamondiferous Kimberlites

Grunsky and Kjarsgaard (2008) evaluated diamond drill core lithogeochemistry from the Star kimberlite located in Saskatchewan, Canada. The Star kimberlite is a series of kimberlite eruptions with five recognized phases (early Joli Fou, mid-Joli Fou, late-Joli Fou, Pense, Cantuar). The early Joli Fou phase contains more macro-diamonds than the other phases, thus making it highly desirable to recognize this phase in the diamond exploration and evaluation process. Figure 1 shows a scatter plot matrix of elements that are typically associated with the kimberlite magma and fractionation processes. The data are presented in cation percent and are untransformed. The scatter plots reveal two distinct patterns. There is a linear pattern of relative enrichment/depletion for all five phases. Additionally, there is a distinct pattern of relative enrichment/depletion associated with the early-, mid- and late-Joli Fou kimberlites. A separate trend is noted for the Pense and Cantuar phases. The late Joli Fou and the Cantuar kimberlites show the greatest relative enrichment of the 5 element shown in this figure. These linear patterns reveal the control that stoichiometry has over the formation of minerals and the associated changes in mineralogy during magma fractionation. Such patterns are well document by Pearce (1968).


Figure 1.Scatter plot matrix of elements that are typically associated with kimberlite magma fractionation from the Star kimberlite geochemistry. Note the linear relationships between these elements. There is one linear trend associated with the evolution of the eJF-mJF-lJF eruption and a second linear trend associated with the Pense and Cantuar eruptions.

The lithogeochemical data were also evaluated in a multivariate process discovery context by the application of a principal component analysis (PCA) on logcentred data. The eigenvalues of the PCA indicated that the first three components accounted for $77 \%$ of the overall data variation, with the first
two accounting for $66 \%$. Figure 2 shows the PCA biplot for the data in which there are three notable compositional trends. The first is a trend toward the positive PC1 and negative PC2 axis in which there is a relative enrichment of $\mathrm{P}, \mathrm{Nb}, \mathrm{La}, \mathrm{Th}$ and Zr that represents kimberlite magma and fractionation processes. Mineral phases associated with this trend include apatite and perovskite. The Pense and Cantuar phases appear to have the highest kimberlite magmatic component, and exhibit more extensive fractionation than the mid- and late-Joli Fou phases. The trend of the mid- and late-Joli Fou phases toward the positive PC1 and PC2 axes represents relative enrichment in $\mathrm{K}, \mathrm{Rb}$ and Na . This represents crustal contamination through the assimilation of feldspar minerals that exist in the upper crust through which the kimberlite magma ascended; alkali feldspars do not crystallize from kimberlite magma. The third trend along the negative PC 1 axis shows relative enrichment in $\mathrm{Si}, \mathrm{Ni}, \mathrm{Mg}, \mathrm{Cr}$, and Co and represents contamination of the kimberlite magma by the earth's lithospheric mantle. These elements are typically associated with the minerals olivine, pyroxene, Cr -spinel and Cr -pyrope garnet. The principal component biplot of the first two elements reveals distinctive information about the processes that have affected the kimberlite geochemical compositions.


Figure 2. PCA biplot of PC1-PC2 for the logcentred Star Kimberlite geochemical data. Three distinct trends are displayed: kimberlite fractionation, crustal contamination and mantle contamination. The screeplot shows that most of the data variation is contained in the first two components.

A linear discriminant analysis was carried out using an additive $\log$ ratio based on Ga as the divisor. Logratio theory demonstrates that the choice of any divisor will yield the same results when using classification methods. In this study, Ga was chosen because it appears to be neutral with respect to the processes observed in the PC1-PC2 biplot, i.e. it plots near the origin on the PC1 - PC2 plot. Table 1 shows the results of the classification using the first three principal components derived from the alr transformed data using Ga as the divisor. The overall accuracy is over $91 \%$ with minimal overlap. This is shown graphically in Figure 3 where the linear discriminant scores are plotted on the first two discriminant axes.


Figure 3. Plot of linear discriminant function scores 1 and 2 for the Star kimberlite geochemical data. The five phases are each enclosed by a convex hull to indicate distinctiveness and overlap. This is shown numerically in Table 1

Table 1. Classification accuracy of the Star kimberlite phases based on linear discriminant analysis of the first seven principal components.

|  | Count Accuracy |  |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: |
|  | Cantuar | eJF | IJF | mJF | Pense |
| Cantuar | 20 | 0 | 0 | 0 | 2 |
| eJF | 0 | 146 | 0 | 4 | 4 |
| IJF | 0 | 0 | 24 | 4 | 0 |
| mJF | 0 | 2 | 1 | 37 | 0 |
| Pense | 1 | 4 | 0 | 0 | 22 |


|  | \% Accuracy |  |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: |
|  | Cantuar | eJF | JJF | mJF | Pense |
| Cantuar | 90.91 | 0.00 | 0.00 | 0.00 | 9.09 |
| eJF | 0.00 | 94.81 | 0.00 | 2.60 | 2.60 |
| IJF | 0.00 | 0.00 | 85.71 | 14.29 | 0.00 |
| mJF | 0.00 | 5.00 | 2.50 | 92.50 | 0.00 |
| Pense | 3.70 | 14.81 | 0.00 | 0.00 | 81.48 |

Overall
Accuracy [\%] 91.88
The effectiveness of process recognition using principal component analysis enables the identification of distinct groups of observations associated with different processes. In this example, the relative proportion of kimberlite magma and crustal and mantle contamination components are identified, in addition to magma fractionation processes. Furthermore, these processes are clearly identified with
specific element associations, which in turn are characteristic of the mineral(s) that are associated with a specific process.

### 3.2 Predictive Mapping Regional Geology Using Lake Sediment Geochemistry

### 3.2.1 Process Discovery

Re-analysis of lake sediments was carried out over three 1:250,000 scale map areas (NTS 65A, 65B, 65C) in southern Nunavut Territory, Canada (McCurdy et al., 2012). The geology of two of the NTS sheets ( $65 \mathrm{~A}, 65 \mathrm{~B}$ ) had been mapped at $1: 250 \mathrm{~K}$ by Eade (1973a, b) and is shown in Figure 4 along with the lake sediment sampling sites. NTS sheet 65 C was compiled at a regional 1:500K scale (Tella et al. 2007) form one million scale maps. No distinction was made by Tella (2007) between two important granitic intrusion types; the Hudson granite ( 1.83 Ga ) and the Nueltin granite ( 1.75 Ga ) (Peterson et al., 2012; 2014).

The three map sheets are located within the southern Hearne Province, a poorly understood terrane located between the central Hearne supracrustal terrane (to the north), which is dominated by ca. 2.7-2.65 Ga mafic-to-felsic, oceanic volcanic rocks and younger tonalite to granite plutons, and the Trans-Hudson orogen (to the south), which forms the northern boundary of the Superior Province. The southern Hearne domain is dominantly comprised of Archean tonalitic and charnokitic gneisses, approximately 2.8 Ga in age. However, strong evidence for fragments of much older crust, up to 3.3 Ga , has been found in the form of inherited Archean zircons and $\mathrm{Sm}-\mathrm{Nd}$ model ages obtained from Proterozoic post-orogenic plutons of the Hudson granite, intruded at about 1.83 Ga . Nueltin rapakivi granite (ca. 1.75 Ga ) is also present in the area. Previously, Proterozoic plutons in this area were distinguished by van Breemen et al. (2005) on the basis of archived hand samples, field descriptions, and partial geochronological data. East of Nueltin Lake, however, identification of individual plutons was hampered by poor exposure and a lower frequency of archival material.


Figure 4. Geology and lake sediment sample sites for NTS 65A/B/C. The geology of NTS65C has not been mapped to include recognition of the Nueltin granite ( $\mathrm{Pp}-\mathrm{Ng}$ ).

Table 2. Dominant map units used for study in NTS 65A/B.

| Legend | \#Sites | Description |
| :---: | :---: | :---: |
|  |  | Archean mafic rocks |
| Ar_mu | 50 | (undifferentiated) |
| Ar_c | 131 | Archean carbonate rocks |
| $\mathrm{Ar}_{-} \mathrm{dt}$ | 266 | Archean diorite |
| Ar_t | 221 | Archean tonalite |
| Ar_s | 111 | Archean sediments |
| Pp_Hgr | 266 | Paleoproterozoic Hudson granite |
| Pp_Hu | 49 | Paleoproterozoic Hurwitz sediments |
| Pp_Nq | 178 | Paleoproterozoic Nueltin Granite |
| Qt | 189 | Quaternary sediments |
| Pp-Wps | 27 | Paleoproterozoic psammitic gneiss |
| Unknown | 833 | Sample sites from NTS 65C |

In this study, the lake sediment geochemistry from two of the map sheets (NTS 65A, 65B) were tagged based on the known bedrock lithology (derived from the detailed geology Eade 1973a, b; Table 2) of the sample site for the purpose of predicting the geology of NTS65C. The lake sediment sample sites from NTS 65C were tagged as "unknown". A suite of 45 elements ( $\mathrm{Ag}, \mathrm{Al}, \mathrm{As}, \mathrm{Au}, \mathrm{Ba}, \mathrm{Be}, \mathrm{Bi}, \mathrm{Ca}, \mathrm{Cd}, \mathrm{Ce}, \mathrm{Co}$, $\mathrm{Cr}, \mathrm{Cs}, \mathrm{Cu}, \mathrm{Fe}, \mathrm{Ga}, \mathrm{Hf}, \mathrm{Hg}, \mathrm{K}, \mathrm{La}, \mathrm{Li}, \mathrm{Mg}, \mathrm{Mn}, \mathrm{Mo}, \mathrm{Na}, \mathrm{Nb}, \mathrm{Ni}, \mathrm{P}, \mathrm{Pb}, \mathrm{Rb}, \mathrm{S}, \mathrm{Sb}, \mathrm{Sc}, \mathrm{Se}, \mathrm{Sn}, \mathrm{Sr}, \mathrm{Th}, \mathrm{Ti}$, Tl, U, V, W, Y, Zn, Zr) were determined by inductively coupled plasma emission spectroscopy/mass spectrometry (ICP-ES/MS) after Aqua regia digestion (McCurdy et al., 2012).
Initially, the data were transformed using the logcentred ratio (Atichison, 1986) from which a principal component analysis was carried out as part of the "discovery process" approach. A tabular summary of those results are shown in Appendix A where the first seven components account for $72.8 \%$ of the overall variation in the data. This is also illustrated in Figure 5, where a "screeplot" of the ordered eigenvalues are shown. The first seven components display a steep decay indicating that these components account for most of the variability of the data. Appendix A also shows R-score values for each element over each principal component. The magnitude and sign of the R -scores indicate the relative significance of a given element with respect to each other for a given component. The significance of the R-scores are directly associated with the magnitude of each eigenvalue. The relative contributions shown in Appendix A indicate the relative significance of an element over the principal components and the absolute contributions indicate the relative significance of an element within a given principal component. Appendix A shows that many of the elements high loadings within the first 5 components. However, some elements such as $\mathrm{Au}, \mathrm{Bi}, \mathrm{Sb}$, and S have a significant amount of variability accounted for in lesser components (PC5, 7, 9, 10) which likely represent significant, but under-sampled processes. These components may be useful vectors for Au mineral exploration follow-up. The content of Appendix A, for the first three principal components is graphically expressed in Figures 6a and 6b.


Figure 5. Ordered eigenvalue plot ("screeplot") of from the principal component analysis of the lake sediment geochemistry.


Figure 6a. Principal component biplot of PC1-PC2 for the lake sediment geochemical data. Convex hulls are shown for the Nueltin granite $(\mathrm{Pp}-\mathrm{Ng})$ and the Hudson granite ( $\mathrm{Pp}-\mathrm{Hgr}$ ).


Figure 6b. Principal component biplot of PC1-PC2 for the lake sediment geochemical data. Only the Nueltin granite ( $\mathrm{Pp}-\mathrm{Ng}$ ) scores are shown in this biplot.


Figure 6c. Principal component biplot of PC1-PC2 for the lake sediment geochemical data. Only the Hudson granite ( $\mathrm{Pp}-\mathrm{Hgr}$ ) scores are shown in this biplot.


Figure 6d. Principal component biplot of PC2-PC3 for the lake sediment geochemical data. Convex hulls for the Nueltin granite $(\mathrm{Pp}-\mathrm{Ng})$ and the Hudson granite $(\mathrm{Pp}-\mathrm{Hgr})$ are shown in this biplot.

Figure 6a shows a principal component biplot of PC1-PC2. The R-scores for each element are located throughout the plot and correspond to the scores shown in Appendix A. The scores of the individual observations (sample sites) are plotted with a symbol and colour that indicates the lithology at the sample site. The legend displays the coding for the sample scores. The two lithologies of interest in this study, the Hudson granite ( $\mathrm{Pp}-\mathrm{Hgr}$ ) and Nueltin granite ( $\mathrm{Pp}-\mathrm{Ng}$ ), are demarcated by ellipses that describe convex hulls for these bedrock geology units. The sample sites from NTS 65c are shown as grey dots throughout the figure. Importantly, many sample sites tagged as Nueltin granite ( $\mathrm{Pp}-\mathrm{Ng}$ ) occur in the negative PC1negative PC2 quadrant and show relative enrichment in Ag-Zn-Mo-Y-La-Ce-Be-P-U. Figures 6a and 6b show biplots of the PC scores for the lake sediment sites coded as Hudson and Nueltin granites. The relative positions of these scores with respect to the principal component loadings of the elements shows that the sample sites associated with the Hudson granite (Figure 6b) show a contrast in relative enrichment of $\mathrm{K}-\mathrm{Ga}-\mathrm{Rb}-\mathrm{Li}-\mathrm{Cs}-\mathrm{Sn}-\mathrm{Th}$ and $\mathrm{Ca}-\mathrm{Sr}-\mathrm{Sb}-\mathrm{Hg}-\mathrm{S}-\mathrm{Ba}-\mathrm{Cu}$. The latter group of elements likely reflects the influence of supracrustal rocks in the lake sediment composition. Figure 6 c shows the PC scores of the lake sediment sample sites coded with the Nueltin granite. The trend of the group of observations in Figure 6c are almost orthogonal to the trend seen in Figure 6b. The relative enrichment trend and contrast for the Nueltin granite sites are $-\mathrm{Ag}-\mathrm{Zn}-\mathrm{Mo}-\mathrm{Y}-\mathrm{La}-\mathrm{Ce}-\mathrm{Be}$ in contrast with $\mathrm{Mg}-\mathrm{Cr}-\mathrm{Sc}-$ Ti-As-Nb-Th-Co-Au. The biplot of PC2-PC3 displayed in Figure 6b shows a distinct trend of relative enrichment of $\mathrm{U},-\mathrm{La}-\mathrm{Ce}-\mathrm{Sn}-\mathrm{Th}-\mathrm{Rb}$, which is associated with the Nueltin granite. This distinction is more clearly illustrated in biplots of Figures 6a, b. Figure $6 b$ shows the biplot with only the pca scores for the Hudson granite and Figure 6c shows the biplot with only the scores of the Nueltin granite. A visual comparison of the two trends suggests that they are nearly orthogonal. In summary, the lake sediment data shows a strong, 1 to 1 correspondence between Nueltin granite and positive lake sediment anomalies in specific incompatible elements, notably REE and Y, in the well-characterized area around Nueltin Lake
Sample sites tagged with the Hudson granite lithology ( $\mathrm{Pp}-\mathrm{Hgr}$ ) occur near the centre of the biplot. The ellipsoid hull that surrounds these points displays a trend from the negative PC 1 - positive PC 2 quadrant ( $\mathrm{Ba}-\mathrm{Cu}-\mathrm{Ca}$ signature) to the positive PC 1 - negative PC 2 quadrant ( $\mathrm{Nb}-\mathrm{Th}-\mathrm{K}-\mathrm{Ga}-\mathrm{Sn}-\mathrm{Cs}-\mathrm{Li}-\mathrm{Rb}$ signature). The association of Archean sediments (Ar-s) in the negative PC1- positive PC2 quadrant of the biplot may be indicative of some mixing of the lithologies in the lake sediments, although this association may also be the result of contamination or partial melting of Archean sediment,. The biplots also show that there is significant overlap of principal component scores for most of the sample sites over a wide range
of lithologies. Thus, it is difficult to define distinct element associations and trends for many of these lithologies.
Following procedures outlined by Grunsky et al. (2014), a thorough geostatistical analysis of the principal component scores was carried out. Modelled semi-variograms and variogram maps (semi-variograms created with an azimuthal range of $0-359^{\circ}$ at $1^{\circ}$ increments) were examined from which kriged images of the principal components were created. Figure 7a shows a variogram map for the first principal component. The image shows a moderate anisotropy with a minimum sill at $75^{\circ}$ east of north and a maximum sill at $145^{\circ}$ east of north. A semi-variogram was modelled at $75^{\circ}$ east of north and shown in Figure 7b. This model was used to generate the kriged image shown in Figure 7c. High values of PC1 represent relative enrichment of $\mathrm{K}, \mathrm{Ti}, \mathrm{Ga}, \mathrm{Rb}, \mathrm{Cs}, \mathrm{Li}, \mathrm{Sn}, \mathrm{Sc}, \mathrm{Mg}$ and Th . This relative enrichment overlies Quaternary sediments (Qt), Archean tonalite (Ar-t) and Archean carbonate (Ar-c) rocks. Negative PC1 scores represent relative enrichment of $\mathrm{Cd}, \mathrm{Se}, \mathrm{S}, \mathrm{Hg}, \mathrm{Ag}$ and overlie a range of Archean lithologies including tonalite and diorite (Ar-t, Ar-dt) and Quaternary sediments (Qt). The lack of direct correspondence of the first principal component with any specific lithology is not unexpected as it represents the dominant geochemical variability that is typical of the major lithologies throughout the area.


Figure 7a. Variogram map of the first principal component derived from the lake sediment geochemical data. The map demonstrates anisotropy trending in a south-easterly direction.


Figure 7b. Modelled semi-variogram of the first principal component derived from the lake sediment geochemical data. The sermi-variogram is derived from an anisotropic search direction with a combined short-range and long-range model.


Figure 7c. Interpolated image of the first principal component using the modelled semi-variogram in Figure 7b. Grid is meters. The range of scores cover most of the dominant rock types and the variability appears to be independent of underlying lithologies.

The variogram map of the second principal component (Figure 8a) shows anisotropy with a low trending at $170^{\circ}$ and a high trending at $45^{\circ}$. A modelled semi-variogram (Figure 8b) modelled at 45 o east of north was used to create the kriged image of Figure 8c. The positive values of PC2 are associated with Archean sediments and carbonate (Ar-s, Ar-c) and the negative values are associated with the Nueltin granite (PpNg ) as is shown in Figure 6a.
Although not shown, the third principal component has positive scores associated with Archean diorite (Ar-dt) and Nueltin granite ( $\mathrm{Pp}-\mathrm{Ng}$ ), and the negative scores have a distinct association with the Hurwitz Group sediments ( $\mathrm{Pp}-\mathrm{Hu}$ ). Positive values of principal component four is distinctly associated with the Nueltin granite and there is no clear association of negative PC4 scores with any specific lithology. The linear combinations of elements that represent variability and association within the metric space defined by PCA is governed solely by the stoichiometry of mineralogy, which is only partly dependent on the underlying lithology. In some cases (PC2, PC3, PC4) there is a clear association with principal component scores and specific lithologies. In other cases (PC1, PC5, PC6) there are no obvious associations between lithology and PC scores. Principal component analysis is an effective method for discovering processes that influence the relationships of the variable within geochemical survey data. Dominant eigenvalues followed by a low rate of decay as exhibited in Figure 5 demonstrate that fact that there is structure in the data. Given this structure the application of classification methods can be applied to validate the structure.


Figure 8a. Variogram map of the second principal component derived from the lake sediment geochemical data.

Nueltin PC2


Figure 8b. Modelled semi-variogram of the second principal component derived from the lake sediment geochemical data. The semi-variogram model is derived from a short range spherical model and a longer range linear model.


Figure 8c. Interpolated image of the second principal component using the modelled semi-variogram in Figure 8 b . Grid is in meters Sites associated with the Nueltin granite show as negative areas.

### 3.2.2 Process Validation

The ability to use lake sediment geochemical data for predictive mapping of underlying lithology requires evaluation in the effectiveness of discriminating between the lithologies. Testing this can be done by applying an Analysis of Variance (AOV) using the logcentred elements or the principal components derived from the logcentred elements. An AOV was carried out on the logcentred data using 10 classes (Ar-c, Ar-dt, Ar-mu, Ar-s, Ar-t, Pp-Hgr, Pp-Hu, Pp-Ng, Pp-Wps, and Pp-Qt) with the elements and another AOV was carried out using the principal components. Figure 9a shows an ordered plot of elements and the associated F -value that provides a measure of class separation. Figure 9a shows that Be-Y-Ni-Ce-La-As-Cr have high F-values and account for much of the lithologic separation. An analysis of variance was also applied to the principal components derived from the logcentred data and Figure 9 b shows a plot of F-value for each principal component. Figure 9 b shows that the second principal component accounts for the majority of lithologic separation followed by PC10, PC14, PC3, PC9, PC18, PC8, PC13, PC7 and PC16. The steep decay of F-values in Figure 9b suggests that the principal component scores are more effective in the application of classification procedures. The advantage of PC scores derived from logcentred data is that they are orthogonal and represent a linear combination of elements that reflect stoichiometry, thereby making them more realistic in the application of classification methods.


Figure 9a. Ordered plot of F-value and elements as a measure for class separation from the logcentred lake sediment geochemical data. Figure 9b. Ordered plot of F-value and principal components as a measure of class separation derived from the logcentred lake sediment geochemical data. The dominance of PC2 over PC1 is reflected in the maps of the principal components. PC2 clearly outlines the two distinct granitic phases (Nueltin and Hudson) whereas these are not recognizable in the map of PC1.

Linear discriminant analysis (LDA) was carried out on ten principal components (PC2, PC10, PC14, PC3, PC9, PC18, PC8, PC13, PC7, and PC16). Cross-validation was used to determine an average accuracy of the classification and the results are shown in Table 3. The first table shows the classification counts for each of the 10 lithologies in a matrix form. The matrix diagonal provides the actual counts that were correctly assigned. The off-diagonal matrix element counts indicate where there is overlap in the classification of the lithologies. These counts are expressed as percentages in the second table. In the second table the classification accuracy for each lithology can be seen along the diagonal and the accuracies range from $28 \%$ ( $\mathrm{Ar}-\mathrm{mu}$ ) to $82 \%$ ( $\mathrm{Ar}-\mathrm{c}$ ). Both the Hudson granite ( $\mathrm{Pp}-\mathrm{Hgr}$ ) and the Nueltin granite ( $\mathrm{Pp}-\mathrm{Ng}$ ) show high prediction accuracy ( $70 \%$ and $77 \%$, respectively). The overall accuracy of the classification is $59 \%$. The off-diagonal matrix elements of the table indicate the percentage overlap or "confusion" in the classification. It is significant to note that there is little overlap between the Hudson and the Nueltin granites, for example $1.1 \% \mathrm{Nu}$ are classified as Hudson, and $0 \%$ Hudson are classified as Nueltin. Both of these granitic rocks show overlap with the Archean tonalite and diorite classes (Ar-t, Ar$\mathrm{dt})$. These relationships are expressed graphically in Figure 10 which is a plot of the linear discriminant scores of the sample site observations that are coded according to the initially assigned lithology (symbol) and the lithology assigned from the classification (colour). It is clear that many of the sites from NTS65C ("unknown") are classified as Nueltin granite (Pp-Ng), Archean tonalite (Ar-t) and sediment (Ar-s). Convex hull ellipses are drawn for each of the 6 of the classes (Pp-Ng, Pp-Hgr, Ar-t, Ar-c, Ar-mu, Ar-s).
Posterior probabilities for each of the lithologies were estimated in the application of the linear discriminant analysis. The probability estimates are based on allocation of assignment for each sample site to at least one of the classes based on the Mahalanobis distance to each class multivariate compositional mean based on the 10 principal components used for the analysis.

|  |  | Ar- | Ar- |  |  | Pp- | Pp- | Pp- |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
|  | Ar-c | dt | mu | Ar-s | Ar-t | Hgr | Hu | Ng | Qt |
| Ar-c | 107 | 11 | 0 | 1 | 0 | 11 | 0 | 0 | 1 |
| Ar-dt | 23 | 127 | 6 | 9 | 39 | 23 | 3 | 5 | 31 |
| Ar-mu | 3 | 10 | 14 | 10 | 5 | 3 | 1 | 0 | 4 |
| Ar-s | 10 | 5 | 4 | 50 | 14 | 13 | 3 | 0 | 12 |
| Ar-t | 7 | 34 | 2 | 4 | 133 | 9 | 5 | 14 | 13 |
| Pp-Hgr | 12 | 24 | 0 | 5 | 8 | 187 | 2 | 0 | 28 |
| Pp-Hu | 0 | 6 | 1 | 5 | 6 | 4 | 22 | 0 | 5 |
| Pp-Ng | 0 | 8 | 0 | 0 | 21 | 2 | 1 | 137 | 9 |
| Qt | 1 | 28 | 4 | 10 | 18 | 32 | 5 | 11 | 80 |
|  |  |  |  |  |  |  |  |  |  |
|  | $\%$ Accuracy |  |  |  |  |  |  |  |  |
|  |  | Ar- | Ar- |  |  | Pp- | Pp- | Pp- |  |
|  | Ar-c | dt | mu | Ar-s | Ar-t | Hgr | Hu | Ng | Qt |
| Ar-c | 81.7 | 8.4 | 0.0 | 0.8 | 0.0 | 8.4 | 0.0 | 0.0 | 0.8 |
| Ar-dt | 8.7 | 47.7 | 2.3 | 3.4 | 14.7 | 8.7 | 1.1 | 1.9 | 11.7 |
| Ar-mu | 6.0 | 20.0 | 28.0 | 20.0 | 10.0 | 6.0 | 2.0 | 0.0 | 8.0 |
| Ar-s | 9.0 | 4.5 | 3.6 | 45.1 | 12.6 | 11.7 | 2.7 | 0.0 | 10.8 |
| Ar-t | 3.2 | 15.4 | 0.9 | 1.8 | 60.2 | 4.1 | 2.3 | 6.3 | 5.9 |
| Pp-Hgr | 4.5 | 9.0 | 0.0 | 1.9 | 3.0 | $\mathbf{7 0 . 3}$ | 0.8 | 0.0 | 10.5 |
| Pp-Hu | 0.0 | 12.2 | 2.0 | 10.2 | 12.2 | 8.2 | 44.9 | 0.0 | 10.2 |
| Pp-Ng | 0.0 | 4.5 | 0.0 | 0.0 | 11.8 | 1.1 | 0.6 | $\mathbf{7 7 . 0}$ | 5.1 |
| Qt | 0.5 | 14.8 | 2.1 | 5.3 | 9.5 | 16.9 | 2.7 | 5.8 | 42.3 |
|  |  |  |  |  |  |  |  |  |  |
| Overall |  |  |  |  |  |  |  |  |  |
| Accuracy |  |  |  |  |  |  |  |  |  |
| (\%) |  | 58.7 |  |  |  |  |  |  |  |

Table 3. Classification accuracy of the lithologies from NTS sheets $65 \mathrm{~A} / \mathrm{B}$ based on linear discriminant analysis of the first seven principal components.


Figure 10. Plot of linear discriminant function scores 1 and 2 for the lake sediment geochemical data based on selected principal components shown in Figure 9b. Convex hulls are enclose the Nueltin and Hudson granite values as defined by their initial assignment.

Variogram maps and modelled semi-variograms (not shown) were created for each of the predicted lithologies. Four of the predicted lithologies are shown in Figure 11a-d. Figure 11a displays a predictive map for Archean carbonate rocks in the form of posterior probabilities. The outline of the Archean carbonate lithologies are shown by the black line surrounding posterior probabilities in the southeast portion of the map. Posterior probabilities greater than 0.5 are shown on the west side of the map in NTS sheet 65 C , where the lithologies are unknown. Figure 11 b shows a posterior probability map for Archean tonalite (Ar-t) that has a strong resemblance to the distribution of tonalite in NTS sheets 65A and 65B. An area of increased posterior probability greater than 0.5 also occurs in NTS sheet 65 C . Figure 11c show a map of posterior probabilities for the Hudson granite ( $\mathrm{Pp}-\mathrm{Hgr}$ ) that closely follows the distribution of the Hudson granite in the east part of the area. The west portion of the map, indicates some indication that the Hudson granite may be present in NTS 65C. Figure 11d shows a map of posterior probability for the Nueltin granite ( $\mathrm{Pp}-\mathrm{Ng}$ ), with high probabilities (greater than 0.8 ) in NTS 65 B , where it closely follows the mapped boundaries, as well as a high probability in the southwest part of NTS 65C. Grunsky et al. (2012b) and Peterson et al. (2014) have verified the predicted presence of this lithology in NTS 65C


Figure 11a. Predictive map of Archean carbonate rocks based on the calculation of posterior probabilities and derived from the linear discriminant analysis shown in Figure 10.
Figure 11b. Predictive map of Archean tonalite rocks based on the calculation of posterior probabilities and derived from the linear discriminant analysis shown in Figure 10.
Figure 11c. Predictive map of Paleoproterozoic Hudson granite based on the calculation of posterior probabilities and derived from the linear discriminant analysis shown in Figure 10.
Figure 11d. Predictive map of Paleoproterozoic Nueltin granite based on the calculation of posterior probabilities and derived from the linear discriminant analysis shown in Figure 10.

## 4 Discussion

The examples presented in this manuscript demonstrate the significance and value of compositional data analysis when applied in conjunction with multivariate statistical procedures and geospatial analysis and presentation. In the case of the Star kimberlite lithogeochemical analyses, the identification of three distinct geochemical trends in the data are illustrated through the combined use of principal component analysis and linear discriminant analysis. Principal components generated from geochemical the data also reflect linear combinations of the elements that are controlled by mineral stoichiometry. This is an important concept in the use of multi-element geochemical data for process discovery and validation. The application of this approach for diamond exploration and mining will result in increased efficiencies in both the exploration of, and beneficiation for diamonds.
The lake sediment geochemical survey data from NTS sheets $65 \mathrm{~A} / \mathrm{B} / \mathrm{C}$ demonstrate that the use of principal component analysis identifies distinct lithologic differences between the two major granitoid suites (Hudson and Nueltin). The use of principal components derived from the logcentred data for use in classification results in more effective discrimination between the classes.
For the lake sediment data, low posterior probabilities can be influenced by significant compositional overlap of the classes, or low initial prior probabilities. Low initial prior probabilities can also indicate a limited geospatial presence in the sampled area. The use of geostatistical procedure describe, model and display the geospatial features of the predicted lithologies is an important part of the process validation approach. As demonstrated in Figures 7 and 8, the use of variogram maps and modelled semi-variograms can be used to define spatially coherent patterns as shown in the kriged images of the first two principal components. Although not shown, the same methodology was applied to generate the kriged images of Figures 11a-d. The prediction of the Nueltin granite with a very high posterior probability, in the southwest part of NTS65C, is sufficient evidence that the geochemical uniqueness and spatial coherence of such patterns are valid.
Other variables aside from principal components might also be used including selected isometric logratio balances or distinct elemental subcompositions. In the use of subcompositions, coherence of the composition should be demonstrated a priori.

## 4 Conclusions

The two examples provided in this study provide evidence that the structure of geochemical data is controlled by mineral stoichiometry and when multi-element geochemical data are used, the structure of the data is revealed. Multivariate statistical methods, such as principal component analysis that was used in this study shows that distinct lithologies can be identified as part of the discover process in evaluating geochemical data in a compositional paradigm. Classification procedures can take advantage of structure in data to yield classifications that are geochemically distinct and geospatially coherent.

## 5 Acknowledgements

The authors would like to thank the organizers of the CoDaWork15 meeting for inviting this contribution. In particular the authors wish to thank Santi Thió-Henestrosa and Josep Antoni Martín Fernández of the University of Girona for their support and guidance.

## 6 References

Aitchison, J., (1986). The statistical analysis of compositional data. Chapman and Hall, New York. 416p.
Aitchison, J. (1999). Logratios and natural laws in compositional data analysis, Mathematical Geology, 31(5), 563-580.

Bivand, R., Pebesma, E., Gomez-Rubio, V., (2013). Applied Spatial Data Analysis with R, 2nd Edition, Springer, 405pp.
Buccianti, A., Mateu-Figueras, G. \& Pawlowsky-Glahn, V., (2006), Compositional Data Analysis in the Geosciences: From Theory to Practice. Geological Society Special Publications 264, 212p.

Eade, K. E., (1973)a. Geology, Nueltin Lake, District of Keewatin, Geological Survey of Canada, Preliminary Map 4-1972, 1973; 1 sheet, doi:10.4095/108984.

Eade, K. E., (1973b). Edehon Lake Area, West Half, District of Keewatin, Geological Survey of Canada, Preliminary Map 3-1972, 1973; 1 sheet, doi:10.4095/108978.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barcelo-Vidal, C., (2003). Isometric Logratio Transformations for Compositional Data Analysis. Mathematical Geology; 35, 279-300.
Garrett, R.G., (1983). Sampling Methodology, Chapter 4, in Statistics and Data Analysis in Geochemical Prospecting, Volume 2, R.J. Howarth (ed.), Handbook of Exploration Geochemistry, G,J.S. Govett (ed.), Elsevier, Amsterdam, 83-110.

Grunsky, E.C., (2001). A program for computing RQ-mode principal components analysis for S-plus and R. Computers \& Geosciences, 27: 229-235.

Grunsky, E.C., (2010), The interpretation of geochemical survey data. Geochemistry: Exploration, Environment and Analysis, 1027-74.
Grunsky, E.C., and Kjarsgaard, B.A., (2008). Classification of eruptive phases of the Star Kimberlite, Saskatchewan, Canada based on statistical treatment of whole-rock geochemical analyses; Applied Geochemistry. v. 23 (12), p. 3321-3336. (ESS Contribution \# 20080330).
Grunsky, E.C. and Bacon-Shone, J., (2011). The stoichiometry of mineral compositions, Proceedings of the 4th International Workshop on Compositional Data Analysis, Sant Feliu de Guixols, Spain.

Grunsky, E.C., Corrigan, D., Mueller, U., Bonham-Carter. G-F., (2012a). Predictive geologic mapping using lake sediment geochemistry in the Melville Peninsula Geological Survey of Canada, Open File 7171, 2012; 1 sheet, doi:10.4095/291901.

Grunsky, E C; McCurdy, M W; Pehrsson, S J; Peterson, T D; Bonham-Carter, G F., (2012b). Predictive geologic mapping and assessing the mineral potential in NTS 65A/B/C, Nunavut, with new regional lake sediment geochemical data; Geological Survey of Canada, Open File 7175, 2012, 1 sheet, doi:10.4095/291920

Hron, K., Templ, M., Filzmoser, P., (2010). Imputation of missing values for compositional data using classical and robust methods, Computational Statistics and Data Analysis, 54 (12), pp. 3095-3107.

Martin-Fernandez, J.A., Barceló-Vidal, C., Pawlowsky-Glahn, V., (2003). Dealing with Zeros and Missing Values in Compositional data Sets Using Nonparametric Imputation, Mathematical Geology, 35(3), 253-278.

McCurdy, M W; McNeil, R J; Day, S J A; Pehrsson, S J., (2012). Regional lake sediment and water geochemical data, Nueltin Lake area, Nunavut (NTS 65A, 65B and 65C), Geological Survey of Canada, Open File 6986, 2012; 13 pages; 1 CD-ROM, doi:10.4095/289888
NUMIN, http://nunavutgeoscience.ca/numin_e.html
Nunavut-Manitoba, Geological Survey of Canada, Open File 5441 (3 sheets), CD-ROM.
Palarea-Albaladejo J, Martin-Fernandez JA., (2008). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. Computers \& Geosciences, 34(8): 902-917

Pawlowsky-Glahn, V. \& Buccianti, A. (eds), (2011). Compositional Data Analysis, Theory and Applications. Wiley \& Sons, New York 378p.
Pearce, T.H., (1968). A contribution to the theory of variation diagrams, Contributions to Mineralogy and Petrology 19, (2), 142 - 157

Pebesma, E.J., (2004). Multivariable geostatistics in S: the gstat package. Computers \& Geosciences, 30: 683-691.

Peterson, T D; Scott, J M J; Jefferson, C W; Tschirhart, V., (2012). Regional potassic alteration corridors spatially related to the 1750 Ma Nueltin Suite in the northeast Thelon Basin region, Nunavut - guides to uranium, gold and silver?, Geological Association of Canada-Mineralogical Association of Canada, Joint Annual Meeting, Programs with Abstracts vol. 35, 2012; p. 1

Peterson, T D; Scott, J M J; Lecheminant, A N; Chorlton, L B; D'Aoust, B M A, (2014). Geological Survey of Canada, Canadian Geoscience Map 158, 2014; 1 sheet, doi:10.4095/293892

R Core Team, (2014). R: A language and environment for statistical computing. $R$ Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rendu, J.-M., (1980). A Case Study: Kriging for Ore Valuation and Mine Planning, Geostatistics, Part 7 of a Series, Engineering and Mining Journal, 181 (1), pp. 114-120.

Scott, J M J; Peterson, T D; McCurdy, M W, (2012). U, Th, REE occurrences within Nueltin granite at Nueltin Lake, Nunavut: recent observations, Geological Survey of Canada, Current Research (Online) no. 2012-1, 15 pages, doi:10.4095/289393

Tella, S; Paul, D; Berman, R G; Davis, W J; Peterson, T D; Pehrsson, S J; Kerswill, J A; Geological Survey of Canada, Open File 5441, 2007; 3 sheets; 1 CD-ROM, doi:10.4095/224573
van Breemen, O., Peterson, T.D., Sandeman, H.A., (2005). U-Pb zircon geochronology and Nd isotope geochemistry of Proterozoic granitoids in the western Churchill Province: intrusive age pattern and Archean source domains, Canadian Journal of Earth Sciences, 42, pp 339-377.

Venables, W.N., Ripley, B.D. (2002). Modern Applied Statistics with S (Fourth Edition). Springer, Berlin 495 p .

Appendix A: Summary of the first seven principal components derived from the logcentred lake sediment geochemistry

|  | Eigenvalues |  | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | $\lambda$ |  | 10.63 | 8.21 | 4.21 | 3.17 | 2.87 | 2.07 | 1.58 |
|  | $\lambda \%$ |  | 23.63 | 18.25 | 9.36 | 7.05 | 6.38 | 4.60 | 3.51 |
|  | $\Sigma \lambda \%$ |  | 23.63 | 41.89 | 51.25 | 58.29 | 64.67 | 69.28 | 72.79 |
| R-Scores |  | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |  |
| Ag |  | -0.6896 | -0.3600 | -0.0652 | -0.4082 | -0.0948 | -0.0645 | 0.0484 |  |
| Al |  | 0.0052 | -0.6932 | -0.2858 | -0.3682 | 0.2255 | -0.0028 | 0.1103 |  |
| As |  | 0.0964 | 0.5020 | -0.3098 | 0.4631 | -0.0386 | 0.1031 | 0.1433 |  |
| Au |  | 0.0156 | 0.2721 | 0.0718 | 0.1991 | -0.4514 | -0.2032 | 0.5678 |  |
| Ba |  | -0.3232 | 0.3548 | -0.2361 | -0.0453 | 0.1789 | 0.3024 | -0.0069 |  |
| Be |  | -0.2746 | -0.7159 | -0.0015 | 0.1360 | 0.2290 | 0.0835 | 0.0305 |  |
| Bi |  | 0.2276 | -0.3891 | 0.0465 | -0.0911 | -0.0228 | 0.0542 | 0.3678 |  |
| Ca |  | -0.4743 | 0.6147 | 0.1971 | -0.0488 | 0.0305 | 0.3787 | -0.1139 |  |
| Cd |  | -0.8030 | -0.0510 | -0.1399 | -0.1633 | 0.0176 | 0.0774 | -0.1275 |  |
| Ce |  | -0.3449 | -0.6510 | 0.2431 | -0.1650 | 0.4263 | -0.1341 | 0.1134 |  |
| Co |  | 0.0241 | 0.2584 | -0.6991 | -0.1210 | 0.1971 | -0.2916 | -0.1694 |  |
| Cr |  | 0.4005 | 0.3444 | -0.5225 | -0.4324 | 0.1579 | -0.0254 | 0.1304 |  |
| Cs |  | 0.6925 | -0.4570 | 0.0460 | -0.2453 | -0.2714 | 0.0779 | -0.0128 |  |
| Cu |  | -0.4470 | 0.1815 | -0.0648 | -0.5171 | -0.2777 | -0.4290 | 0.0366 |  |
| Fe |  | 0.0820 | -0.1493 | -0.5839 | 0.3819 | 0.4700 | -0.1869 | -0.1537 |  |
| Ga |  | 0.7319 | -0.2796 | -0.1433 | -0.4141 | 0.1087 | 0.1728 | 0.0884 |  |
| Hf |  | 0.3365 | 0.6999 | 0.3523 | 0.0863 | 0.2250 | -0.1225 | -0.1581 |  |
| Hg |  | -0.5653 | 0.4059 | 0.0649 | -0.2984 | 0.0992 | 0.0415 | -0.0768 |  |
| K |  | 0.8470 | -0.1086 | -0.0075 | -0.2016 | -0.2658 | 0.1034 | -0.1063 |  |
| La |  | -0.3857 | -0.6293 | 0.3271 | -0.1658 | 0.3479 | -0.1474 | 0.1056 |  |
| Li |  | 0.7502 | -0.4464 | 0.0044 | -0.2162 | -0.2853 | 0.0690 | -0.0848 |  |
| Mg |  | 0.6934 | 0.4307 | -0.1744 | -0.1494 | -0.1714 | 0.2361 | -0.0460 |  |
| Mn |  | -0.0738 | -0.0796 | -0.5872 | 0.5283 | 0.0159 | 0.1597 | -0.0231 |  |
| Mo |  | -0.4525 | -0.5731 | 0.0191 | 0.2718 | -0.1175 | -0.0892 | -0.3368 |  |
| Na |  | 0.0236 | 0.0595 | 0.1551 | -0.2829 | -0.2194 | 0.4713 | -0.4132 |  |
| Nb |  | 0.4089 | -0.0611 | 0.4859 | -0.3059 | 0.3879 | 0.0059 | -0.1595 |  |
| Ni |  | -0.0469 | 0.4945 | -0.4123 | -0.4900 | -0.0553 | -0.3776 | 0.0277 |  |
| P |  | -0.2586 | -0.4128 | -0.5709 | 0.0324 | 0.0914 | 0.4426 | 0.1062 |  |
| Pb |  | 0.0266 | -0.0242 | 0.2412 | -0.1052 | 0.5382 | 0.2861 | 0.2982 |  |
| Rb |  | 0.8060 | -0.3750 | 0.1331 | -0.2019 | -0.2013 | 0.0807 | -0.0855 |  |
| S |  | -0.6413 | 0.3283 | 0.2189 | -0.2093 | -0.1775 | -0.0567 | -0.3736 |  |
| Sb |  | -0.3693 | 0.5554 | 0.0347 | -0.0039 | -0.1051 | 0.2420 | 0.0315 |  |
| Sc |  | 0.6690 | 0.1889 | -0.0053 | 0.1187 | 0.3886 | -0.2306 | -0.2066 |  |
| Se |  | -0.7685 | 0.0353 | 0.0363 | -0.1865 | -0.0536 | 0.0686 | -0.1227 |  |
| Sn |  | 0.6104 | -0.4045 | 0.2635 | -0.0518 | 0.1928 | 0.3464 | -0.0125 |  |
| Sr |  | -0.3994 | 0.5417 | 0.1022 | -0.2750 | 0.4482 | 0.2338 | 0.0405 |  |
| Th |  | 0.5716 | -0.0714 | 0.3269 | 0.4279 | 0.3280 | -0.3448 | -0.1567 |  |


| Ti | 0.9094 | 0.1670 | 0.0073 | -0.1135 | 0.1189 | 0.0252 | -0.1138 |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| TI | 0.3195 | -0.5839 | -0.1892 | -0.0986 | -0.2403 | -0.2824 | -0.2146 |
| U | -0.2758 | -0.4211 | 0.4373 | 0.2342 | -0.3026 | -0.0621 | -0.0397 |
| V | 0.3048 | -0.0219 | -0.6501 | -0.1289 | 0.3147 | 0.0346 | -0.0835 |
| W | -0.1031 | -0.4562 | -0.1866 | 0.3853 | -0.1297 | 0.2744 | -0.2798 |
| Y | -0.4994 | -0.6529 | 0.2535 | -0.0544 | 0.2755 | -0.0378 | 0.0049 |
| Zn | -0.4385 | -0.4681 | -0.3566 | -0.1699 | -0.0125 | -0.0915 | -0.2943 |
| Zr | 0.3283 | 0.7017 | 0.4047 | 0.0199 | $\mathbf{0 . 2 6 7 0}$ | -0.1594 | -0.1600 |


| Relative Contributions |  |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
| Ag | 47.5809 | 12.9652 | 0.4254 | 16.6696 | 0.8983 | 0.4168 | 0.2346 |
| Al | 0.0027 | 48.0709 | 8.1692 | 13.5636 | 5.0854 | 0.0008 | 1.2181 |
| As | 0.9304 | 25.2093 | 9.6036 | 21.4559 | 0.1488 | 1.0626 | 2.0554 |
| Au | 0.0243 | 7.4055 | 0.5161 | 3.9643 | 20.3829 | 4.1312 | 32.2588 |
| Ba | 10.4501 | 12.5913 | 5.5757 | 0.2049 | 3.2023 | 9.1515 | 0.0047 |
| Be | 7.5460 | 51.2669 | 0.0002 | 1.8513 | 5.2475 | 0.6982 | 0.0930 |
| Bi | 5.1816 | 15.1480 | 0.2164 | 0.8298 | 0.0520 | 0.2944 | 13.5354 |
| Ca | 22.5056 | 37.8040 | 3.8882 | 0.2382 | 0.0930 | 14.3466 | 1.2989 |
| Cd | 64.5047 | 0.2603 | 1.9584 | 2.6671 | 0.0309 | 0.5992 | 1.6258 |
| Ce | 11.8979 | 42.4009 | 5.9103 | 2.7225 | 18.1834 | 1.7998 | 1.2867 |
| Co | 0.0582 | 6.6789 | 48.8894 | 1.4643 | 3.8847 | 8.5044 | 2.8721 |
| Cr | 16.0471 | 11.8655 | 27.3116 | 18.7029 | 2.4933 | 0.0643 | 1.7021 |
| Cs | 47.9793 | 20.8972 | 0.2121 | 6.0197 | 7.3714 | 0.6069 | 0.0165 |
| Cu | 19.9867 | 3.2940 | 0.4201 | 26.7503 | 7.7156 | 18.4129 | 0.1338 |
| Fe | 0.6720 | 2.2290 | 34.1040 | 14.5916 | 22.0995 | 3.4942 | 2.3621 |
| Ga | 53.5866 | 7.8207 | 2.0547 | 17.1519 | 1.1811 | 2.9856 | 0.7809 |
| Hf | 11.3276 | 49.0005 | 12.4140 | 0.7444 | 5.0637 | 1.5008 | 2.5013 |
| Hg | 31.9687 | 16.4852 | 0.4211 | 8.9087 | 0.9847 | 0.1727 | 0.5906 |
| K | 71.7801 | 1.1798 | 0.0056 | 4.0661 | 7.0695 | 1.0704 | 1.1295 |
| La | 14.8804 | 39.6189 | 10.7020 | 2.7486 | 12.1102 | 2.1722 | 1.1165 |
| Li | 56.3047 | 19.9365 | 0.0019 | 4.6778 | 8.1438 | 0.4759 | 0.7200 |
| Mg | 48.1002 | 18.5566 | 3.0432 | 2.2334 | 2.9397 | 5.5783 | 0.2119 |
| Mn | 0.5443 | 0.6341 | 34.5010 | 27.9216 | 0.0254 | 2.5506 | 0.0534 |
| Mo | 20.4828 | 32.8594 | 0.0366 | 7.3922 | 1.3819 | 0.7954 | 11.3483 |
| Na | 0.0558 | 0.3543 | 2.4078 | 8.0043 | 4.8150 | 22.2186 | 17.0774 |
| Nb | 16.7311 | 0.3733 | 23.6240 | 9.3641 | 15.0570 | 0.0035 | 2.5462 |
| Ni | 0.2201 | 24.4641 | 17.0040 | 24.0195 | 0.3059 | 14.2657 | 0.0767 |
| P | 6.6900 | 17.0471 | 32.6053 | 0.1050 | 0.8358 | 19.5941 | 1.1274 |
| Pb | 0.0710 | 0.0588 | 5.8193 | 1.1079 | 28.9736 | 8.1893 | 8.8933 |
| Rb | 64.9861 | 14.0714 | 1.7719 | 4.0761 | 4.0526 | 0.6510 | 0.7311 |
| S | 41.1495 | 10.7839 | 4.7924 | 4.3827 | 3.1523 | 0.3215 | 13.9660 |
| Sb | 13.6422 | 30.8653 | 0.1203 | 0.0016 | 1.1053 | 5.8572 | 0.0996 |
| Sc | 44.7749 | 3.5709 | 0.0028 | 1.4102 | 15.1108 | 5.3222 | 4.2716 |
| Se | 59.0890 | 0.1249 | 0.1319 | 3.4799 | 0.2870 | 0.4708 | 1.5067 |
| Sn | 37.2800 | 16.3731 | 6.9460 | 0.2681 | 3.7187 | 12.0059 | 0.0157 |


| Sr | 15.9626 | 29.3557 | 1.0451 | 7.5633 | 20.0953 | 5.4695 | 0.1639 |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| Th | 32.6887 | 0.5099 | 10.6924 | 18.3150 | 10.7648 | 11.8954 | 2.4556 |
| Ti | 82.7357 | 2.7906 | 0.0053 | 1.2893 | 1.4152 | 0.0637 | 1.2961 |
| Tl | 10.2129 | 34.1041 | 3.5828 | 0.9716 | 5.7767 | 7.9797 | 4.6067 |
| U | 7.6113 | 17.7408 | 19.1341 | 5.4896 | 9.1582 | 0.3852 | 0.1575 |
| V | 9.2929 | 0.0481 | 42.2819 | 1.6611 | 9.9078 | 0.1199 | 0.6981 |
| W | 1.0627 | 20.8222 | 3.4840 | 14.8545 | 1.6826 | 7.5351 | 7.8332 |
| Y | 24.9539 | 42.6497 | 6.4278 | 0.2956 | 7.5932 | 0.1426 | 0.0024 |
| Zn | 19.2395 | 21.9176 | 12.7235 | 2.8888 | 0.0156 | 0.8381 | 8.6661 |
| Zr | 10.7850 | 49.2607 | 16.3838 | 0.0398 | 7.1294 | 2.5410 | 2.5620 |

Absolute Contributions

|  | PC1 | PC2 | PC3 | PC4 | PC5 | PC6 | PC7 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Ag | 4.4737 | 1.5783 | 0.1010 | 5.2564 | 0.3133 | 0.2016 | 0.1486 |
| AI | 0.0003 | 5.8518 | 1.9387 | 4.2770 | 1.7735 | 0.0004 | 0.7714 |
| As | 0.0875 | 3.0688 | 2.2791 | 6.7657 | 0.0519 | 0.5139 | 1.3017 |
| Au | 0.0023 | 0.9015 | 0.1225 | 1.2500 | 7.1085 | 1.9981 | 20.4294 |
| Ba | 0.9825 | 1.5328 | 1.3232 | 0.0646 | 1.1168 | 4.4262 | 0.0030 |
| Be | 0.7095 | 6.2409 | 0.0001 | 0.5838 | 1.8300 | 0.3377 | 0.0589 |
| Bi | 0.4872 | 1.8440 | 0.0514 | 0.2617 | 0.0181 | 0.1424 | 8.5719 |
| Ca | 2.1160 | 4.6020 | 0.9228 | 0.0751 | 0.0324 | 6.9389 | 0.8226 |
| Cd | 6.0649 | 0.0317 | 0.4648 | 0.8410 | 0.0108 | 0.2898 | 1.0296 |
| Ce | 1.1187 | 5.1616 | 1.4027 | 0.8585 | 6.3414 | 0.8705 | 0.8149 |
| Co | 0.0055 | 0.8130 | 11.6026 | 0.4617 | 1.3548 | 4.1133 | 1.8189 |
| Cr | 1.5088 | 1.4444 | 6.4817 | 5.8976 | 0.8695 | 0.0311 | 1.0779 |
| Cs | 4.5111 | 2.5439 | 0.0503 | 1.8982 | 2.5707 | 0.2935 | 0.0105 |
| Cu | 1.8792 | 0.4010 | 0.0997 | 8.4352 | 2.6908 | 8.9056 | 0.0847 |
| Fe | 0.0632 | 0.2713 | 8.0937 | 4.6012 | 7.7071 | 1.6900 | 1.4959 |
| Ga | 5.0383 | 0.9520 | 0.4876 | 5.4085 | 0.4119 | 1.4440 | 0.4946 |
| Hf | 1.0650 | 5.9650 | 2.9461 | 0.2347 | 1.7660 | 0.7259 | 1.5841 |
| Hg | 3.0058 | 2.0068 | 0.0999 | 2.8092 | 0.3434 | 0.0835 | 0.3740 |
| K | 6.7489 | 0.1436 | 0.0013 | 1.2822 | 2.4654 | 0.5177 | 0.7153 |
| La | 1.3991 | 4.8230 | 2.5398 | 0.8667 | 4.2234 | 1.0506 | 0.7071 |
| Li | 5.2939 | 2.4269 | 0.0005 | 1.4750 | 2.8401 | 0.2302 | 0.4560 |
| M g | 4.5225 | 2.2590 | 0.7222 | 0.7042 | 1.0252 | 2.6980 | 0.1342 |
| Mn | 0.0512 | 0.0772 | 8.1879 | 8.8045 | 0.0089 | 1.2336 | 0.0338 |
| Mo | 1.9258 | 4.0001 | 0.0087 | 2.3310 | 0.4819 | 0.3847 | 7.1868 |
| Na | 0.0053 | 0.0431 | 0.5714 | 2.5240 | 1.6792 | 10.7463 | 10.8151 |
| Nb | 1.5731 | 0.0454 | 5.6065 | 2.9528 | 5.2511 | 0.0017 | 1.6125 |
| Ni | 0.0207 | 2.9781 | 4.0354 | 7.5740 | 0.1067 | 6.8998 | 0.0486 |
| P | 0.6290 | 2.0752 | 7.7380 | 0.0331 | 0.2915 | 9.4769 | 0.7140 |
| Pb | 0.0067 | 0.0072 | 1.3811 | 0.3493 | 10.1044 | 3.9609 | 5.6321 |
| Rb | 6.1101 | 1.7130 | 0.4205 | 1.2853 | 1.4133 | 0.3149 | 0.4630 |
| S | 3.8690 | 1.3128 | 1.1374 | 1.3820 | 1.0994 | 0.1555 | 8.8446 |
| Sb | 1.2827 | 3.7573 | 0.0285 | 0.0005 | 0.3855 | 2.8329 | 0.0631 |
| Sc | 4.2098 | 0.4347 | 0.0007 | 0.4447 | 5.2698 | 2.5741 | 2.7052 |


| Se | 5.5557 | 0.0152 | 0.0313 | 1.0973 | 0.1001 | 0.2277 | 0.9542 |
| :--- | ---: | ---: | ---: | ---: | :--- | :--- | :--- |
| Sn | 3.5052 | 1.9932 | 1.6484 | 0.0845 | 1.2969 | 5.8068 | 0.0100 |
| Sr | 1.5008 | 3.5736 | 0.2480 | 2.3849 | 7.0082 | 2.6454 | 0.1038 |
| Th | 3.0735 | 0.0621 | 2.5375 | 5.7753 | 3.7542 | 5.7533 | 1.5551 |
| Ti | 7.7790 | 0.3397 | 0.0012 | 0.4066 | 0.4935 | 0.0308 | 0.8208 |
| TI | 0.9602 | 4.1516 | 0.8503 | 0.3064 | 2.0146 | 3.8595 | 2.9174 |
| U | 0.7156 | 2.1597 | 4.5410 | 1.7310 | 3.1939 | 0.1863 | 0.0998 |
| V | 0.8737 | 0.0059 | 10.0345 | 0.5238 | 3.4553 | 0.0580 | 0.4421 |
| W | 0.0999 | 2.5348 | 0.8268 | 4.6841 | 0.5868 | 3.6444 | 4.9607 |
| Y | 2.3462 | 5.1919 | 1.5255 | 0.0932 | 2.6481 | 0.0690 | 0.0015 |
| Zn | 1.8089 | 2.6681 | 3.0196 | 0.9109 | 0.0054 | 0.4053 | 5.4882 |
| Zr | 1.0140 | 5.9967 | 3.8882 | 0.0125 | 2.4864 | 1.2290 | 1.6225 |

# Detecting and mapping compositional global outliers to identify mineral exploration targets, Case study: Khusf district, East of Iran 

Majid Keykha Hosseinpoor ${ }^{1}$, Hamid Moini ${ }^{2}$, and Farhad Mohammad Torab ${ }^{3}$<br>${ }^{1}$ PhD student, Yazd University, Yazd, Iran; majidkeykha@yahoo.com<br>${ }^{2} \mathrm{PhD}$ student, Yazd University, Yazd, Iran<br>${ }^{3}$ Assistant professor , Yazd University, Yazd, Iran


#### Abstract

To identify mineral promising targets in khusf area located in the east of Iran, a geochemical stream sediment study was done and 652 samples were analyzed for 23 elements. Although there are some different known classical convenient methods to delineate the univariate anomalies, the most reliable one is to deal with the geochemical anomalies through compositional global multivariate outliers that on this case study has been addressed. After preparation of the raw data (missing and censored values imputation on compositional data using zCompositions package in $\mathbf{R}$ ), the global outliers of the data were detected and marked in the coordinated map of the samples using mvoutlier package in $\mathbf{R}$. Robust principal components of the compositional data were biplotted as well. The biplots showed the elements affecting the global outliers. Compositional univariate outlier plot is another useful tool in the package that was used to visualize each outlying observation position element-wise. Comparing the resulted global outliers with the officially reported mineralization indexes and field observations confirmed that mvoutlier package is the fastest and most reliable tool amongst the other convenient ways of locating geochemical anomalies provided that the sampling and data preparation stages are done properly.


Key words: Geochemical data, stream sediments, logratio transformation, global outlier, PCA, Khusf.

## 1 Introduction

The processing of geochemical data for detecting multivariate geochemical patterns or signals associated with mineralization in support of mineral resource exploration is challenging (Zuo, 2014). Identifying and detection of anomaly area is an important issue in stream sediment geochemical exploration studies (Stanley and Sinclair, 1989; Stanley and Sinclair, 1991; Ghavami-Riabi and others, 2010). Different methods for calculation of threshold values in either uni- or multi-element approaches have been developed in many investigations (Matchullat and others, 2000; Reimann and others, 2005). As an important source of geo-information, geochemical datasets of multiple element concentrations have been successfully processed by advanced multivariate analytical methods (e.g., factor analysis, principal component analysis, cluster analysis, etc.) to identify mineralizationassociated geological bodies and delineate mineralization-favored spaces (Bogoch and others, 1993; Brantley and White, 2009; Cheng, 2007; Cheng and others, 2011; Hao and others, 2007; Yousefi and others, 2012, Yousefi and others, 2014, Wang and others, 2011, 2012, 2013; Zhao and others, 2012, 2013, 2014). Outlier detection as one of the main tasks in multivariate data analysis is challenging and if preparing a dataset (especially in extreme values, which may not be extreme in any of the individual attribute space directions) is correctly done and there is no pollution in study area, most likely it can represent the geochemical anomaly areas (Filzmoser and others, 2005, 2012).

In the convenient geoscience researches, exploratory datasets are often compositional data which have long been of concern in the geochemical field (Aitchison, 1982, 1984; Buccianti, 2011; Buccianti and others, 2006; Carranza, 2011; Chayes and Trochimczyk, 1978; Rollinson, 1992). A stream sediment geochemical dataset is an example of a closed number system because it contains compositional variables that are parts of a whole (Carranza, 2011). Classical approaches in geochemical data analysis are mainly univariate and applying normal or lognormal transformations after outlier detection in that space that most of the times proposes the removal of the outliers. In the last five decades, several researchers have discussed the problems in statistical analysis of closed number systems such as compositional datasets (Aitchison, 1981, 1983, 1984, 1986, 1999; Aitchison and others, 2000; Buccianti and others, 2005; Chayes, 1960; Egozcue and others, 2003, 2005; Filzmoser and others, 2009; Miesch, 1969; Thió-Henestrosa and others, 2005). The statistical analysis of compositional multivariate data is a much discussed topic in the field of multivariate statistics (Filzmoser and others, 2009). In practice, log-ratio transformations are commonly employed in geochemical data processing to open closed systems for better understanding of realistic relationships among compositions (Filzmoser and others, 2012; Gallo and others, 2013; Verma and others, 2006; ).

In present case study to identify exploratory target areas in khusf district located in east of Iran, we used compositional robust PCA together with adaptive multivariate outlier detection method to delineate and map the compositional global outliers. A recent method and tool developed by Peter Filzmoser and other scientists (Filzmoser and Hron, 2007; Filzmoser and others, 2011; Filzmoser and others, 2005; Filzmoser and Geschwandtner, 2011)

## 2 Compositional or closed data

The geochemical data especially stream sediments analysis has an intrinsic compositional properties. They are multivariate observations that describe quantitatively the parts of some whole. Thus, their components carry exclusively relative information about the parts (Aitchison, 1986). Typically these observations are expressed as data with a constant sum constraint such as proportions, percentages, or $\mathrm{mg} / \mathrm{kg}$ (Filzmoser and Hron, 2007), i.e:

$$
\begin{equation*}
X=\left(x_{1}, \ldots, x_{D}\right)^{\prime}, x_{i}>0, \sum_{i=1}^{D} x_{i}=\text { constant } \tag{1}
\end{equation*}
$$

The D-dimensional space defined in Equation (1) is called a Simplex. Due to the fact that geochemical data follow the simplex geometry, classical statistical methods that rely mostly on the euclidean geometry produce spurious results when they are applied to raw compositional data (Filzmoser and Hron, 2008; Filzmoser and others, 2009; Hron and others, 2010). Whether or not the data follow a normal distribution is of no importance at all (Filzmoser and others, 2012). In order to deal with such data that arose challenges in scientific literatures ,Aitchison (1986) and Egozcue (2003) introduced three transformations to open and bring the data to real euclidean space and classical statistics could analyze them. They are called additive logratio (alr), centered logratio ( $c l r$ ) and isometric logratio transformation (ilr). Of all these transformations the two latter are considerable because they show an isometric relationship between real and simplex spaces. Compositions $X \in S^{D}$ are transformed via clr to data $Y \in R^{D}$, with

$$
\begin{equation*}
Y=\operatorname{clr}(X)=\left(y_{1}, \ldots, y_{D}\right)^{\prime}=\left(\ln \frac{x_{1}}{\sqrt[D]{\prod_{i=1}^{D} x_{i}}}, \ldots, \ln \frac{x_{D}}{\sqrt[D]{\prod_{i=1}^{D} x_{i}}}\right)^{\prime} \tag{2}
\end{equation*}
$$

But this transformation results in collinear data because $\sum_{i=1}^{D} y_{i}=0$.
The other transformation is ilr that solves the problem of data collinearity resulting from the clr transformation, while preserving all its advantageous properties (Egozcue and others, 2003). Compositions $X \in S^{D}$ are transformed based on the choice of an orthonormal basis to data $Z \in R^{D-1}$, with

$$
\begin{equation*}
Z=i \operatorname{lr}(X)=\left(z_{1}, \ldots, z_{D-1}\right)^{\prime}, z_{i}=\sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_{j}}}{x_{i+1}} \tag{3}
\end{equation*}
$$

One of the biggest problems in applying this transformation is that it would be complicated to interpret the results because of the reduction in dimension, especially in geochemistry that the dimensions are the analyzed elements. One way is to apply ilr in multivariate analysis and then backtransform the results in $c l r$ so that the results could be meaningful.

Due to the definition of compositional data, all the relevant information about $x_{1}$ is contained in the ratios to each of the remaining parts $x_{2}, \ldots, x_{D}$. Accordingly, this relative information for all remaining parts needs to be considered also for univariate data analysis (Filzmoser and others, 2009). The ilr transformation from Equation (3) can be used for this purpose. The ilr variable:

$$
\begin{equation*}
z_{1}=\sqrt{\frac{D-1}{D}} \ln \frac{x_{1}}{\sqrt[D-1]{\prod_{j=2}^{D} x_{j}}} \tag{4}
\end{equation*}
$$

contains all the relative information between $x_{1}$ and $x_{2}, \ldots, x_{D}$, because none of $z_{2}, \ldots, z_{D-1}$ includes $x_{1}$. In this way, each compositional part can be expressed by a single ilr variable as defined in Equation (4) which here is used for univariate analysis of global outliers (Filzmoser and others, 2009).

For all logratio transformations, the problem of missing or zero values should be solved prior to any analysis. Fortunately there have been proposed many solutions and tools to impute them in the best way. One of them is zCompositions package written and developed by Javier PalareaAlbaladejo and Josep Antoni Martín-Fernández that provides principled methods to deal with zeros and nondetects in compositional data sets (Palarea-Albaladejo and Martín-Fernández, 2015).

### 2.1 Global outliers detection

In contrast to univariate outliers, multivariate outliers are not necessarily extreme along single coordinates. Rather, they could deviate from the multivariate data structure formed by the majority of observations (Filzmoser and others, 2005). The estimated covariance structure is used to assign a distance to each observation indicating how far the observation is from the center of the data cloud with respect to the covariance structure (Hron and others, 2010; Filzmoser and Hron, 2011).

This distance measure is the well-known Mahalanobis distance, defined for a sample $x_{1}, \ldots, x_{n}$ of $n$ observations in the $d$-dimensional real space $R^{D}$ as

$$
\begin{equation*}
M D\left(x_{i}\right)=\left[\left(x_{i}-T\right)^{\prime} C^{-1}\left(x_{i}-T\right)\right]^{1 / 2} \quad \text { for } i=1, \ldots, n \tag{5}
\end{equation*}
$$

T and C in Equation (5) are location and spread estimators, respectively. In the case of multivariate normally distributed data, the arithmetic mean and the sample covariance matrix are the best choices, leading to the best statistical efficiency (Filzmoser and others, 2012). In this case, the squared Mahalanobis distances approximate a chi-square distribution $\chi_{d}^{2}$ with $d$ degrees of freedom. A certain cut-off value like the $97.5 \%$ quantile of $\chi_{d}^{2}$ can be taken as an indication of extremeness: data points with higher (squared) Mahalanobis distance than the cut-off value are considered as potential outliers (Rousseeuw and Van Zomeren, 1990).
Both the arithmetic mean and the sample covariance matrix are highly sensitive to outlying observations (Maronna and others, 2006). A number of robust estimators of covariance have been proposed in the literature, like the $\mathrm{MCD}^{1}$ estimator (Maronna and others, 2006). It looks for a subset $h$ out of $n$ observations with the smallest determinant of their sample covariance matrix. The subset size $h$ can vary between half the sample size and $n$, and it will determine not only the robustness of the estimates, but also their efficiency. It is common to use the same cut-off value from the $\chi_{d}^{2}$ distribution (Rousseeuw and Van Zomeren, 1990). Filzmoser (2005) introduced a more advanced approach to the cut-off value that could lead to more accurate values. This method called "adaptive outlier detection", accounts for the actual numbers of observations and variables in the data set, and it tries to distinguish among extremes of the data distribution and outliers coming from a different distribution (Filzmoser and others, 2005).

As with the other multivariate methods applied to compositional data, it is important to use an appropriate data transformation first. Either the clr transformation or a proper choice of the ilr transformation can be used for this purpose (Filzmoser and Hron, 2008; Filzmoser and others, 2012).

Several methods have been proposed for the identification of multivariate outliers, making use of robust statistics (Maronna and others, 2006). Such tools have also been developed in the context of compositional data (Filzmoser and Hron, 2008). The package mvoutlier in $\mathbf{R}$ software presented by Filzmoser and Gschwandtner (2011) is a free and precious tool to deal with outliers especially in compositional data. It is designed on a comprehensive mathematical basis that detects, maps and plots the global outliers in a multivariate dataset. This is done by computing for each observation the distances to the medians along the single ilr variables. The median of all distances determines the color (or gray scale): a high value, resulting in a red (or dark) symbol, means that most univariate parts have higher values than the average, and a low value, resulting in a blue (or light) symbol, refers to an observation with mainly low values. This characterization helps to interpret multivariate outliers (Filzmoser and others, 2012). The robust squared Mahalanobis distances are computed and split by four values: the quantiles $0.25,0.5$, and 0.75 and the aforementioned adaptive outlier cutoff. By default, the symbols for the resulting five groups (in the above order) are large open circle, small open circle, point, small "+" and large "+" (Filzmoser and others, 2012). The user can replace the symbols with observations (samples) numbers to see the outlying samples. In case of mapping the outliers it requires the coordinates of taken samples to locate them on the sampling area.

### 2.2 Geochemical anomalies

In the last step of exploratory geochemical data, it is convenient to identify promising targets known as anomalies on a map. There has been proposed many methods to delineate the univariate anomalies i.e. defining thresholds with mean $+2 S D$ or median $+2 M A D$ that both require normal distribution assumption (Carranza, 2008). Although these methods have been used a lot in literatures, they produce unreliable results because of the compositional nature of the geochemical

[^2]data. The best way is robust multivariate analysis of this kind of data to investigate not only the relations but also the enriched elemental associations.

Multivariate outliers often show atypical phenomena and mostly come from different populations. In geochemical stream sediments that are taken for mineralization prospecting, they show interesting anomalies indicating mineralized zones up or down a catchment basin. Using mvoutlier package a geochemist can detect the samples that are potential multivariate outliers (global or local) and if it is superimposed on a geological map, it can mark the anomalies because of which these samples are identified as being outliers.

On the other hand, to check the anomalies of global outliers, we used the mineralized indexes marked in the local geological map produced by geological survey of $\operatorname{Iran}^{2}$ and its related official reports and also direct field check.


Figure 1: (a) Structural zones of Iran, and (b) simplified geological map of the study area.

## 3 Study area and geochemical data

The study area measures $2500 \mathrm{~km}^{2}$ and is covered by the $1: 100,000$ scale quadrangle map of khusf prepared by GSI (Vahdati, 1989). The study area is a small region in the northern part of the central lut block in east of Iran and lies between longitudes E58 $30^{\prime}$ and $\mathrm{E} 59^{\circ}$ and $\mathrm{N} 32^{\circ} 30^{\prime}$ and $\mathrm{N} 33^{\circ}$ latitudes. Due to its location in the northern part of the central lut block, this area has inherited arid conditions (Arjmandzadeh and others, 2011). The lut block extends over 900 km in a north-south direction and is only 200 km wide in an east-west direction (Stocklin and others, 1973). It is confined by the nayband fault and shotori range on the west (Berberian and others, 1982). The magmatic activity in the lut block began in the middle jurassic 165-162 Ma and reached

[^3]its peak in the tertiary (Karimpour, 2011). Volcanic and subvolcanic rocks of tertiary age cover over half of lut block with up to $2000 m$ thickness and were formed due to subduction prior to the collision of the arabian and asian plates (Camp and others, 1982; Tirrul and others, 1983; Berberian, 2000).

There are widespread exposures of late cretaceous-early tertiary sedimentary rocks and cenozoic volcanism. Elevated areas and mountain ranges are arranged in the north and north-eastern part, while in the other parts, the topography is dominated by abundant irregular hills and intervening alluvial plains within which scattered, higher and isolated volcanic bodies may be seen (Vahdati, 1989).

The oldest rocks in the area are upper paleozoic which are restricted to small fractured and faulted fragments of central iranian paleozoic platform which are exposed as an antiform in the northwest of the map. The lithology units of the study area are flysch type sediments, marl, limestone and mid-eocene, non-volcanic deposits have been found in the northwest and southwest corner of the sheet including agglomerate, ignimbrite, marl, tuff and eocene-oligocene volcanic rocks including andesite, tuff, dacitic andesite and finally include eocene-oligocene dacite, silicified volcanic rocks, oligo-miocene rhyolites and dacites. The simplified lithological map of the study area is shown in


Figure 2: The stream sediment samples location and $\mathrm{Fe}, \mathrm{Pb}, \mathrm{Cu}$ indexes in the study area.

Figure 1. To identify promising exploratory areas in khusf $1: 100,000$ sheet, a drainage geochemical survey was carried out by GSI and 652 geochemical samples were taken. Figure 2 shows the stream sediment samples location in the study area. It also shows the major mineralization indexes extracted from GSI reports in the area. Some in the form of veinlets or outcrops and typical minerals. They are important in validating the accuracy of analysis. The minus 80 -mesh fraction of the stream sediments was analyzed for 20 elements including $\mathrm{Au}, \mathrm{W}, \mathrm{Mo}, \mathrm{Zn}, \mathrm{Pb}, \mathrm{Ag}, \mathrm{Cr}, \mathrm{Ni}$, $\mathrm{Bi}, \mathrm{Sc}, \mathrm{Cu}, \mathrm{As}, \mathrm{Sb}, \mathrm{Cd}, \mathrm{Co}, \mathrm{Sn}, \mathrm{Ba}, \mathrm{V}, \mathrm{Sr}$ and Hg , and three oxides: $\mathrm{MnO}, \mathrm{TiO}_{2}$ and $\mathrm{Fe}_{2} \mathrm{O}_{3}$.


Figure 3: The global outliers location marked with (a)only outliers and (b)all observations in the study area. (Red colored symbols show the considerable outliers that represent probable geochemical anomalies).

## 4 Discussion

First of all $\mathrm{Au}, \mathrm{W}$, Mo and three oxides were removed from the dataset prior to analysis. The first three because of their problems and the oxides for homogenizing the data. The final dataset is a matrix of 652 samples (or observations) by 17 elements (or variables). Then the most important thing that should be done before any data analysis is preparing the dataset by imputation of missing and zero values provided that they are a small fraction of the variable (in this study they did not exceed $10 \%$ ). This was done with the valuable $\mathbf{R}$ software packages "compositions" and "zCompositions" that using detection limits vector replace the 'bad' values with the proper ones within delicate algorithms considering compositional properties of the data.

Finally the data was compositionally scaled and global outliers were detected and mapped (Fig. 3). The PCA biplots (Fig. 4) and univariate outliers plot (Fig. 5) also were drawn.

It can be seen in PCA biplot that PC1 \& PC2 explain about $83 \%$ variability and almost all the outliers are from $\mathrm{Cr}, \mathrm{Co}, \mathrm{Ni}, \mathrm{Cu}, \mathrm{Pb}, \mathrm{Zn}$ elements. These are the major lithological constituents that are the host rocks and sources of mineralizations in the area. Comparing the results of univariate ilr-elements with the map shows that samples $177,178,147,149$ located in the northeastern part of the area are outliers for $\mathrm{Pb}, \mathrm{As}, \mathrm{Ag}, \mathrm{Cd}, \mathrm{Bi} \& \mathrm{Sr}$. The samples comply with Pb mineralization indexes upstream that basin. Samples 22, 25, 26, 27, 28, 30, 31, 34, 36, 37, 39, 40 located in the eastern part of the map are mostly enriched in Cu and together with samples $53,55,54,61$ in the southeastern part that are high in $\mathrm{Ni}, \mathrm{Zn} \& \mathrm{Bi}$ form a principal component of the data. The indexes of Cu mineralization in the form of cuprite and malachite which were observed upstream that basin, confirmed the anomalous outliers. Samples $39 \& 40$ in the center show high Cr and Co, highlighted as another principal component. Samples $82,127,181$ in the northern central part are high in Sc \& V, the two elements that match with the Fe-index and explain rich Fe-oxides blanket covering parts of the area. Individual samples $15 \& 16$ also show high Co and Cr and low Cu .


Figure 4: The PCA biplots marked with (a)all observations and (b)only outliers. (Red colored symbols just like (Fig. 3) show the considerable outliers).


Figure 5: The univariate outlier plot without the regular observations.

## 5 Conclusion

In this study we applied a robust multivariate outlier detection method on a brownfields exploration area in east of Iran. High conformity between obtained results by this method and known deposit indexes in the study area is the main advantage of using this method.

If the sampling of stream sediments is carried out in order to find the promising areas (and not for the environmental purposes) and samples are taken far from other pollution sources, the outliers can be representatives of just the mineralization anomalies because they are abnormal populations emerging out of the normal background of the study area. Otherwise it will be every extreme reason such as pollution, error or a combination of them and in that case it needs an extra field
check to distinguish the mineralization from other processes. Therefore the interpretation would become complicated.

Although this study was done on a geochemically known area and many researches have been done by far on this area, this was just to evaluate and confirm the mentioned method. The authors have tested and applied this approach on many other similar projects in Iran and outstanding target anomalies were detected.

## Acknowledgements

we would like to express our gratitude to the authorities of south Khorasan Industries, Mines and Trade Organization who provided access to the data. We also appreciate dear doctor Mohammad Shiva, the assistant professor of Birjand university and the director of this project, for his kind comments, points and recommendations and the help at hand through the study.

## References

Aitchison, J. (1981). A new approach to null correlations of proportions. Journal of the International Association for Mathematical Geology, 13(2), 175-189.

Aitchison, J. (1982). The statistical analysis of compositional data. J. R. Stat. Soc. Ser. B Methodol., 44, 139-177.

Aitchison, J. (1983). Principal component analysis of compositional data. Biometrika, 70(1), 57-65.

Aitchison, J. (1984). The statistical analysis of geochemical compositions. Journal of the International Association for Mathematical Geology, 16(6), 531-564.

Aitchison, J. (1986). The statistical analysis of compositional data, (Vol. 25): Chapman \& Hall.
Aitchison, J. (1999). Logratios and natural laws in compositional data analysis. Mathematical Geology, 31(5), 563-580.

Aitchison, J., Barceló-Vidal, C., Martin-Fernández, J., \& Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. Mathematical Geology, 32(3), 271-275.

Arjmandzadeh, R., Karimpour, M., Mazaheri, S., Santos, J., Medina, J., \& Homam, S. (2011). SrNd isotope geochemistry and petrogenesis of the Chah-Shaljami granitoids (Lut Block, Eastern Iran). Journal of Asian Earth Sciences, 41(3),283-296.

Berberian, F., Muir, I.D., Pankhurst, R.J., Berberian, M., (1982). Late cretaceous and early miocene andean type plutonic activity in northern Makran and central Iran. Journal of the Geological Society, 139, 605- 614.

Berberian, M., (2000). The 1994 Sefidabeh earthquakes in eastern Iran: Blind thrusting-plane slip on growing anticline, \& active tectonics of the Sistan suture zone. Geophys. J. Int., 142, 283-299.

Bogoch, R., Shirav, M., Beyth, M., Halicz, L. (1993). Geochemistry of ephemeral stream sediments in the Precambrian mountainous arid terrain of southern Israel. Journal of Geochemical Exploration, 46, 349-364.

Brantley, S.L., White, A.F. (2009). Approaches to modeling weathered regolith. Rev. Mineral. Geochem., 70, 435-484.

Buccianti, A., \& Pawlowsky-Glahn, V. (2005). New perspectives on water chemistry and compo-
sitional data analysis. Mathematical Geology, 37(7), 703-727.
Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (2006). Compositional data analysis in the geosciences: from theory to practice. Geol. Soc. Lond. Spec. Publ. 264, 212.

Buccianti, A. (2011). Natural laws governing the distribution of the elements in geochemistry: the role of the log-ratio approach. In: Pawlowsky-Glahn, V., Buccianti, A. (Eds.), Compositional Data Analysis: Theory and Applications, Chapter 18, pp. 255--266.

Camp, V., Griffis R., (1982). Character, genesis and tectonic setting of igneous rocks in the Sistan suture zone, eastern Iran. Lithous, 15, 221-239.

Carranza, E.J.M. (2008). Geochemical anomaly and mineral prospectivity mapping in GIS, (Vol.11): Elsevier

Carranza, E.J.M. (2011). Analysis and mapping of geochemical anomalies using logratio-transformed stream sediment data with censored values. Journal of Geochemical Exploration, 110(2), 167185.

Chayes, F. (1960). On correlation between variables of constant sum. Journal of Geophysical Research, 65(12), 4185-4193.
Chayes, F., Trochimczyk, J. (1978). An effect of closure on the structure of principal components. Math. Geol., 10, 323-333.

Cheng, Q. (2007). Mapping singularities with stream sediment geochemical data for prediction of undiscovered mineral deposits in Gejiu, Yunnan Province, China. Ore Geol. Rev., 32, 314-324.

Cheng, Q., Bonham-Carter, G., Wang, W., Zhang, S., Li, W., Xia, Q. (2011). A spatially weighted principal component analysis for multi-element geochemical data for mapping locations of felsic intrusions in the Gejiu mineral district of Yunnan, China. Comput. Geosci., 37, 662-669.

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

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

Filzmoser, P., Garrett, R.G., Reimann, C. (2005). Multivariate outlier detection in exploration geochemistry. Computers $\mathcal{E}^{\mathcal{G}}$ Geosciences, 31(3), pp. 579-585.

Filzmoser, P., Hron, K. (2008). Outlier detection for compositional data using robust methods. Math. Geosci.. 40, 233-248.

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

Filzmoser, P., Hron, K., \& Reimann, C. (2009). Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Science of the Total Environment, 407(23), 6100-6108.

Filzmoser, P., Geschwandtner, M. (2011). mvoutlier: Multivariate outlier detection based on robust methods. Manual and package, version 1.9.1. http://cran.r-project.org/package= mvoutlier

Filzmoser, P., Hron, K., Reimann, C. (2012). Interpretation of multivariate outliers for compositional data. Computers $\mathfrak{B}^{3}$ Geosciences, 39, pp. 77-85.

Gallo, M., \& Buccianti, A. (2013). Weighted principal component analysis for compositional data: application example for the water chemistry of the Arno River (Tuscany, central Italy). Environmetrics, 24(4), 269-277.

Ghavami-Riabi, R., Seyedrahimi-Niaraq, M. M., Khalokakaie, R., \& Hazareh, M. R. (2010). Uspatial statistic data modeled on a probability diagram for investigation of mineralization phases and exploration of shear zone gold deposits. Journal of Geochemical Exploration, 104(1), 27-33.

Hao, L., Lu, J., Li, L., Mo, G., Yan, G., Shi, Y., Zhao, Y. (2007). Method of using regional geochemical data in geological mapping in shallow overburden areas. Geol. China, 34, 710-715

Hron, K., Templ, M., Filzmoser, P. (2010). Imputation of missing values for compositional data using classical and robust methods. Computational Statistics and Data Analysis, 54 (12), 3095-3107.

Karimpour, M. (2011). Review of age, Rb-Sr geochemistry and petrogenesis of Jurassic to Quaternary igneous rocks in Lut Block, Eastern Iran. Geopersia, 1(1), 19-54.

Maronna, R., Martin, R.D., Yohai, V.J. (2006). Robust Statistics: Theory and Methods. Wiley, New York.

Matchullat, J., Ottenstein, R., Reimann, C. (2000). Geochemical background - can we calculate it? Environmental Geology, 39, 990-1000.

Miesch, A. (1969). The constant sum problem in geochemistry Computer applications in the earth sciences, (pp. 161-176): Springer.
Palarea-Albaladejo, J., \& Martín-Fernández, J. A. (2015). zCompositions-R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics and Intelligent Laboratory Systems, 143, 85-96.

Reimann, C., Filzmoser, P., \& Garrett, R. G. (2002). Factor analysis applied to regional geochemical data: problems and possibilities. Applied geochemistry, 17(3), 185-206.

Reimann, C., Filzmoser, P., Garrett, R.G. (2005). Background and threshold: critical comparsion of methods of determination. Sci. Total Environ., 346, 1-16.

Rollinson, H. (1992). Another look at the constant sum problem in geochemistry. Mineral. Mag., 56, 469-475.

Rousseeuw, P.J., Van Zomeren, B.C. (1990). Unmasking multivariate outliers and leverage points. Journal of the American Statistical Association, 85(411), pp. 633-651.

R Development Core Team (2011). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Stanley, C.R., Sinclair, A.J. (1989). Comparison of probability plots and the gap statistic in selection of threshold of exploration geochemistry data. Journal of Geochemical Exploration, 32, 355-357.

Stanley, C.R., Sinclair, A.J. (1991). A fundamental approach to threshold estimation in exploration geochemistry, probability plots revisited. Journal of Geochemical Exploration. 41, 1-22.

Stocklin, J., Nabavi, M.H., (1973). Tectonic map of Iran. Geological Survey of Iran publication.
Thió-Henestrosa, S., \& Martín-Fernández, J. (2005). Dealing with compositional data: the freeware CoDaPack. Mathematical Geology, 37(7), 773-793.

Tirrul, R., Bell, I.R., Griffis, R.J., Camp, V.E., (1983). The Sistan suture zone of eastern iran. Geological Society of America Bulletin, 94(1), 134-156.

Vahdati, F., (1989). Geological Map of khusf (1:100,000), Geological Survey of Iran publication.
Verma, S. P., Guevara, M., \& Agrawal, S. (2006). Discriminating four tectonic settings: Five new geochemical diagrams for basic and ultrabasic volcanic rocks based on log-ratio transformation of major-element data. Journal of Earth System Science, 115(5), 485-528.

Wang, W., Zhao, J., Cheng, Q. (2011). Analysis and integration of geo-information to identify granitic intrusions as exploration targets in southeastern Yunnan District, China. Comp. Geosci., 37, 1946-1957.

Wang, W., Zhao, J., Cheng, Q., Liu, J. (2012). Tectonic-geochemical exploration modeling for characterizing geo-anomalies in southeastern Yunnan district, China. Journal of Geochemical Exploration, 122, 71-80.

Wang, W., Zhao, J., Cheng, Q. (2013). Fault trace-oriented singularity mapping technique to characterize anisotropic geochemical signatures in Gejiu mineral district, China. Journal of Geochemical Exploration, 134, 27-37.

Yousefi, M., Kamkar-Rouhani, A., \& Carranza, E.J.M. (2012). Geochemical mineralization probability index (GMPI): a new approach to generate enhanced stream sediment geochemical evidential map for increasing probability of success in mineral potential mapping. Journal of Geochemical Exploration, 115, 24-35.

Yousefi, M., Kamkar-Rouhani, A., \& Carranza, E.J.M. (2014). Application of staged factor analysis and logistic function to create a fuzzy stream sediment geochemical evidence layer for mineral prospectivity mapping. Geochemistry: Exploration, Environment, Analysis, 14(1), 45-58.

Zhao, J., Wang, W., Dong, L., Yang, W., Cheng, Q. (2012). Application of geochemical anomaly identification methods in mapping of intermediate and felsic igneous rocks in eastern Tianshan, China. Journal of Geochemical Exploration, 122, 81-89.

Zhao, J., Wang, W., Cheng, Q. (2013). Investigation of spatially non-stationary influences oftectono-magmatic processes on Fe mineralization in eastern Tianshan, China with geographically weighted regression. Journal of Geochemical Exploration, 134, 38-50.

Zhao, J., Wang, W., Cheng, Q. (2014). Application of geographically weighted regression to identify spatially non-stationary relationships between Fe mineralization and its controlling factors in eastern Tianshan, China. Ore Geol. Rev., 57, 628-638.

Zuo, R. (2014). Identification of geochemical anomalies associated with mineralization in the Fanshan district, Fujian, China. Journal of Geochemical Exploration, 139, 170-176.

# Space-time compositional models: an introduction to simplicial partial differential operators 

JARAUTA-BRAGULAT ${ }^{1}$, E. and EGOZCUE ${ }^{1}$, J. J.<br>${ }^{1}$ Dept. Matemàtica Aplicada III - ETS Enginyers de Camins, Canals i Ports de Barcelona<br>Universitat Politècnica de Catalunya, Barcelona, Spain<br>(e-mail: eusebi.jarauta@upc.edu ; juan.jose.egozcue@upc.edu)


#### Abstract

A function assigning a composition to space-time points is called a compositional or simplicial field. These fields can be analysed using the compositional analysis tools. A study of linear models for evolutionary compositions depending on one variable, usually time, was formulated by Egozcue and Jarauta-Bragulat (2014) in terms of the so-called simplicial linear differential equations. The foundations of differential and integral calculus for simplex-valued functions of one real variable, was presented by Egozcue, Jarauta-Bragulat and Díaz-Barrero (2011).


In order to study compositions depending on space and/or time, reformulation and interpretation of traditional partial differential operators is required. These operators such as: partial derivatives, compositional gradient, directional derivative and divergence are of primary importance to state alternative models of processes as diffusion, advection and waves, from the compositional perspective. This kind of models, usually based on continuity of mass, circulation of a vector field along a curve and flux through surfaces, should be analyzed when compositional operators are used instead of the traditional gradient or divergence. This study is aimed at setting up the definitions, mathematical basis and interpretation of such operators.

## 1 Introduction

In a large number of processes studied in Sciences and in Engineering, magnitudes or variables involved can be modelled by a vector. This vector may be a function of one or several variables. Furthermore, in many cases the studied vector is a composition. Study of the evolutionary compositions depending on one variable was introduced by Egozcue and Jarauta-Bragulat (2014). The present work focuses on compositions whose evolution depends on several variables; in many cases, these variables are spatial coordinates and time. For example, the study of the evolution of a pollutant carried by a fluid stream.

For spatial coordinates, a subset $S \in \mathbb{R}^{d}$ of a $d$-dimensional real space is considered and identified with a physical domain; consequently $d=1, d=2$ and $d=3$ are the common choices of that dimension. A location in this space is denoted $s \in S$ and represented in a Cartesian coordinate system; for $d=1$ the spatial coordinate is usually denoted $s=x$; for $d=2, s=(x, y)$ and so on. For time, a subset $T \subseteq \mathbb{R}$ is considered and a point is denoted $t \in T$. For spatial and time evolutionary processes, a domain $S \times T \subseteq \mathbb{R}^{d} \times \mathbb{R} \quad(d=1,2,3)$ is considered.

A space-time vector-valued field with positive components $\mathbf{Z}$ is a function $\mathbf{Z}: S \times T \subseteq \mathbb{R}^{d} \times \mathbb{R} \rightarrow \mathbb{R}_{+}^{n}$ that assigns a positive component vector $\mathbf{Z}(s, t) \in \mathbb{R}_{+}^{n}$ to a space-time point $(s, t) \in S \times T$. If closure operation is then applied: $\mathcal{C} \mathbf{Z}(s, t)=\mathbf{z}(s, t)$ a space-time simplicial field $\mathbf{z}$ (STSF) is obtained. Consequently, a STSF is a function $\mathbf{z}: S \times T \subseteq \mathbb{R}^{d} \times \mathbb{R} \rightarrow \mathbb{S}^{n}$, that assigns a composition $\mathbf{z}(s, t)$ in the $n$-part simplex $\mathbb{S}^{n}$ to any space-time point $(s, t) \in S \times T$. In general, differentiability of $\mathbf{z}(s, t)$ up to the second order is assumed, thus guaranteeing the existence of continuous derivatives up to second order.

## 2 Derivatives and Integrals of a Space-Time Simplicial Field

In the following, definitions and properties are developed for $d=2$ and extension to other values of $d$ are natural. Based on the definition of (ordinary) simplicial derivative (Egozcue et al., 2011) and the real calculus, natural definitions for partial simplicial derivatives of a STSF follow.

Definition 2.1 (Spatial and time simplicial derivatives) Let $\mathbf{z}: S \times T \subseteq \mathbb{R}^{2} \times \mathbb{R} \rightarrow \mathbb{S}^{n}$ be $a$ STSF and $(x, y) \in S$. The spatial-partial simplicial derivatives of $\mathbf{z}$ are

$$
\begin{aligned}
& \partial_{x}^{\oplus} \mathbf{z}(x, y, t)=\lim _{h \rightarrow 0}\left(\frac{1}{h} \odot(\mathbf{z}(x+h, y, t) \ominus \mathbf{z}(x, y, t))\right), \\
& \partial_{y}^{\oplus} \mathbf{z}(x, y, t)=\lim _{h \rightarrow 0}\left(\frac{1}{h} \odot(\mathbf{z}(x, y+h, t) \ominus \mathbf{z}(x, y, t))\right) .
\end{aligned}
$$

The time-partial simplicial derivative is

$$
\partial_{t}^{\oplus} \mathbf{z}(x, y, t)=\lim _{h \rightarrow 0}\left(\frac{1}{h} \odot(\mathbf{z}(x, y, t+h) \ominus \mathbf{z}(x, y, t))\right)
$$

DEFINITION 2.2 (DIRECTIONAL SIMPLICIAL DERIVATIVES) Let $\mathbf{z}: S \times T \subseteq \mathbb{R}^{2} \times \mathbb{R} \rightarrow \mathbb{S}^{n}$ be $a$ STSF. Let $(x, y) \in S$ and $\vec{u}=\left(u_{x}, u_{y}\right)$ be a vector in $\mathbb{R}^{2}$. The simplicial derivative of $\mathbf{z}$ with respect to $\vec{u}$ is

$$
\partial_{\vec{u}}^{\oplus} \mathbf{z}(x, y, t)=\lim _{h \rightarrow 0}\left(\frac{1}{h} \odot\left(\mathbf{z}\left(x+h u_{x}, y+h u_{y}, t\right) \ominus \mathbf{z}(x, y, t)\right)\right)
$$

If $\vec{u}$ is a unit vector, the derivative is called the directional simplicial derivative of $\mathbf{z}$.

Spatial-partial and time-partial simplicial derivatives are computed as they were ordinary simplicial derivatives of a single variable simplex-valued function as developed in Egozcue et al. (2011). Consequently, they can be computed as ordinary derivatives of the log-transformation of the STSF, and then transformed back into compositions. The same scheme works for clr and ilr transformations of $\mathbf{z}$ (Aitchison, 1986; Egozcue et al., 2003). The following proposition summarizes this kind of computation.

Proposition 2.1 Partial simplicial derivatives of $\mathbf{z}(x, y, t)$ can be computed as

$$
\partial_{x}^{\oplus} \mathbf{z}(x, y, t)=\mathcal{C} \exp \left(\partial_{x} \log (\mathbf{z}(x, y, t))\right)=\mathcal{C} \exp \left(\begin{array}{c}
\frac{\partial_{x} z_{1}(x, y, t)}{z_{1}(x, y, t)} \\
\vdots \\
\frac{\partial_{x} z_{n}(x, y, t)}{z_{n}(x, y, t)}
\end{array}\right)
$$

Additionaly

$$
\operatorname{clr}\left(\partial_{x}^{\oplus} \mathbf{z}(x, y, t)\right)=\partial_{x} \operatorname{clr}(\mathbf{z}(x, y, t)) ; \quad \operatorname{ilr}\left(\partial_{x}^{\oplus} \mathbf{z}(x, y, t)\right)=\partial_{x} \operatorname{ilr}(\mathbf{z}(x, y, t))
$$

Similar expressions hold for $\partial_{y}^{\oplus} \mathbf{z}(x, y, t)$ and $\partial_{t}^{\oplus} \mathbf{z}(x, y, t)$.

Definitions and properties related to integral of simplex-valued functions of one variable have been stated in (Egozcue et al., 2011). They are summarized in the next Proposition.

Proposition 2.2 Let $\mathbf{f}: I \subseteq \mathbb{R} \rightarrow \mathbb{S}^{n}$ be a continuous simplex-valued function of real variable.

- A differentiable function $\mathbf{F}: I \subseteq \mathbb{R} \rightarrow \mathbb{S}^{n}$ is an antiderivative of $\mathbf{f}$ on $I$ if, and only if, $\partial^{\oplus} \mathbf{F}(\xi)=\mathbf{f}(\xi), \xi \in I$.
- $\int{ }^{\oplus} d \xi \odot \mathbf{f}(\xi)=\mathcal{C} \exp \left(\int \log (\mathbf{f}(\xi)) d \xi\right)$.
- $\int_{[a, b]}^{\oplus} d \xi \odot \mathbf{f}(\xi)=\mathcal{C} \exp \left(\int_{a}^{b} \log (\mathbf{f}(\xi)) d \xi\right)$.
- $\operatorname{clr}\left(\int_{[a, b]}^{\oplus} d \xi \odot \mathbf{f}(\xi)\right)=\int_{a}^{b} \operatorname{clr}(\mathbf{f}(\xi)) d \xi$.
- $\operatorname{ilr}\left(\int_{[a, b]}^{\oplus} d \xi \odot \mathbf{f}(\xi)\right)=\int_{a}^{b} \operatorname{ilr}(\mathbf{f}(\xi)) d \xi$.

Some natural extended definitions and properties can be stated for double and line integrals.

Definition 2.3 (Double integrals) Let $\mathbf{z}: S \times T \subseteq \mathbb{R}^{2} \times \mathbb{R} \rightarrow \mathbb{S}^{n}$ be a STSF. The double integral of $\mathbf{z}$ in $S$ is a composition in $\mathbb{S}^{n}$ given by

$$
\iint_{S}^{\oplus}(d \xi d \eta) \odot \mathbf{z}(\xi, \eta, t)=\mathcal{C} \exp \left(\iint_{S} \log (\mathbf{z}(\xi, \eta, t)) d \xi d \eta\right)
$$

Proposition 2.3 Properties of double integrals are:

$$
\begin{aligned}
\operatorname{clr}\left(\iint_{S}^{\oplus}(d \xi d \eta) \odot \mathbf{z}(\xi, \eta, t)\right) & =\iint_{S} \operatorname{clr}(\mathbf{z}(\xi, \eta, t)) d \xi d \eta \\
\operatorname{ilr}\left(\iint_{S}^{\oplus}(d \xi d \eta) \odot \mathbf{z}(\xi, \eta, t)\right) & =\iint_{S} \operatorname{ilr}(\mathbf{z}(\xi, \eta, t)) d \xi d \eta
\end{aligned}
$$

Definition 2.4 (Line integrals) Let $\mathbf{z}: S \times T \subseteq \mathbb{R}^{2} \times \mathbb{R} \rightarrow \mathbb{S}^{n}$ be a STSF. The line integral of $\mathbf{z}$ along a regular curve $\Gamma \subset S$ of finite length, is

$$
\int_{\Gamma}^{\oplus} d s \odot \mathbf{z}(x(u), y(u), t)=\mathcal{C} \exp \left(\int_{a}^{b} \log (\mathbf{z}(x(u), y(u), t)) s^{\prime}(u) d u\right)
$$

where $\gamma_{1}=(x(a), y(a)), \gamma_{2}=(x(b), y(b))$ are the end points of the curve $\Gamma$ and $s^{\prime}(u)$ denotes the ordinary derivative of $s$ with respect to the parameter $u$.

Proposition 2.4 Properties of line integrals are:

$$
\begin{aligned}
& \operatorname{clr}\left(\int_{\Gamma}^{\oplus}(d s \odot \mathbf{z}(x(u), y(u), t))\right)=\int_{a}^{b} \operatorname{clr}(\mathbf{z}(x(u), y(u), t)) s^{\prime}(u) d u \\
& \operatorname{ilr}\left(\int_{\Gamma}^{\oplus}(d s \odot \mathbf{z}(x(u), y(u), t))\right)=\int_{a}^{b} \operatorname{ilr}(\mathbf{z}(x(u), y(u), t)) s^{\prime}(u) d u
\end{aligned}
$$

## 3 Simplicial Mass Continuity Equation and Differential Operators

When $n$ species are mixed in a continuum, a common assumption is that the mass of each species changes according the input-output of it through the border of a fixed control volume $V$. When working in a space of dimension $d=2$, volume means area, or alternatively, for $d=1$ is just a length. Notation $V$ is used both for referring to the volume itself and for indicating its magnitude in
some volume unit. The continuity of mass for each species is normally described by the continuity equation (Landau and Lifshitz, 1987; White, 1991). It can be written as

$$
\begin{equation*}
\partial_{t} \rho_{k}+\operatorname{div}\left(\rho_{k} \vec{v}_{k}\right)=0 \quad, \quad k=1,2, \ldots n, \tag{1}
\end{equation*}
$$

where $\rho_{k}$ is the mass density of the $k$-species, and $\vec{v}_{k}=\left(v_{k x}, v_{k y}\right)$ is its velocity in a planar movement. Attention should be paid to the definition of $\rho_{k}$. It is the ratio of the mass $m_{k}$ to the volume $V_{k}$ occupied by the mass of the $k$-species. Accordingly, $\rho_{k}=m_{k} / V_{k}$ and the units can be, for instance, $\mathrm{g} / \mathrm{cm}^{3}$. Interest is centered in the behaviour of (mass) concentration $c_{k}=m_{k} / M$ of each species, which is given as the ratio of the mass of the $k$-species to the total mass $M$ within some given control volume $V$. The overall density is $\rho=M / V$ which leads to a convenient expression of $\rho_{k}$ in terms of concentrations

$$
\begin{equation*}
\rho_{k}=\frac{m_{k}}{V_{k}}=\frac{m_{k} / M}{V_{k} / \rho V}=\rho \frac{m_{k} / M}{V_{k} / V}=\rho \frac{c_{k}}{a_{k}}=\rho d_{k}, k=1,2, \ldots, n \tag{2}
\end{equation*}
$$

where $a_{k}=V_{k} / V$ is the volume fraction or volume concentration of the $k$-species. The ratio of mass to volume concentration is denoted $d_{k}=c_{k} / a_{k}$. Note that the continuity equations hold for each species but not for the total, as the change of concentrations modifies the mass content of $V$ and the diffusion or selective transport of some species may change the overall density $\rho$. As a conclusion continuity equation (1) does not hold for the overall density $\rho$ and different velocities $\vec{v}_{k}$. For a planar flow the fields of densities are considered functions of space location $s \in S \subseteq \mathbb{R}^{2}$ and time $t \in T \subseteq \mathbb{R}$. The explicit dependence is suppressed unless it is necessary, for instance, $\rho(s, t)$ is denoted $\rho$.
Substituting in Eq. (1) the expression of $\rho_{k}$ in Eq. (2) and developing the divergence

$$
\begin{equation*}
\partial_{t}\left(\rho d_{k}\right)+\operatorname{div}\left(\rho d_{k} \vec{v}_{k}\right)=\partial_{t}\left(\rho d_{k}\right)+\partial_{x}\left(\rho d_{k} v_{k x}\right)+\partial_{y}\left(\rho d_{k} v_{k y}\right)=0 \quad, \quad k=1,2, \ldots n \tag{3}
\end{equation*}
$$

In order to introduce logarithmic derivatives, the equation is divided by $\rho d_{k}$

$$
\begin{equation*}
\frac{\partial_{t}\left(\rho d_{k}\right)}{\rho d_{k}}+\frac{\partial_{x}\left(\rho d_{k} v_{k x}\right)}{\rho d_{k}}+\frac{\partial_{y}\left(\rho d_{k} v_{k y}\right)}{\rho d_{k}}=0 \quad, \quad k=1,2, \ldots n . \tag{4}
\end{equation*}
$$

This equation is transformed into

$$
\begin{align*}
\partial_{t} \log \left(\rho d_{k}\right) & =-\left(v_{k x} \partial_{x} \log \left(\rho d_{k}\right)+v_{k y} \partial_{y} \log \left(\rho d_{k}\right)\right)-\left(\partial_{x} v_{k x}+\partial_{y} v_{k y}\right) \\
& =-\left\langle\vec{v}_{k}, \operatorname{grad} \log \left(\rho d_{k}\right)\right\rangle-\operatorname{div}\left(\vec{v}_{k}\right) \\
& =-\partial_{\vec{v}_{k}} \log \left(\rho d_{k}\right)-\operatorname{div}\left(\vec{v}_{k}\right), \quad k=1,2, \ldots, n \tag{5}
\end{align*}
$$

where $\langle\cdot, \cdot\rangle$ is the standard Euclidean inner product in $\mathbb{R}^{2}$ and known properties of derivatives of functions of several variables have been applied. Equations (5) for $k=1,2, \ldots, n$ can be placed in an array as

$$
\left(\begin{array}{c}
\partial_{t} \log \left(\rho d_{1}\right)  \tag{6}\\
\partial_{t} \log \left(\rho d_{2}\right) \\
\vdots \\
\partial_{t} \log \left(\rho d_{n}\right)
\end{array}\right)=\left(\begin{array}{c}
-\partial_{\vec{v}_{1}} \log \left(\rho d_{1}\right) \\
-\partial_{\vec{v}_{1}} \log \left(\rho d_{2}\right) \\
\vdots \\
-\partial_{\vec{v}_{1}} \log \left(\rho d_{n}\right)
\end{array}\right)+\left(\begin{array}{c}
-\operatorname{div}\left(\vec{v}_{1}\right) \\
-\operatorname{div}\left(\vec{v}_{2}\right) \\
\vdots \\
-\operatorname{div}\left(\vec{v}_{n}\right)
\end{array}\right)
$$

Logarithmic derivatives in left-hand side of Equation (6) are transformed into a simplicial derivative by taking $\operatorname{clr}^{-1}$

$$
\mathcal{C} \exp \left(\begin{array}{c}
\partial_{t} \log \left(\rho d_{1}\right)  \tag{7}\\
\partial_{t} \log \left(\rho d_{2}\right) \\
\vdots \\
\partial_{t} \log \left(\rho d_{n}\right)
\end{array}\right)=\mathcal{C} \exp \left(\begin{array}{c}
\partial_{t} \log \rho \\
\partial_{t} \log \rho \\
\vdots \\
\partial_{t} \log \rho
\end{array}\right) \oplus \mathcal{C} \exp \left(\begin{array}{c}
\partial_{t} \log d_{1} \\
\partial_{t} \log d_{2} \\
\vdots \\
\partial_{t} \log d_{n}
\end{array}\right)=\partial_{t}^{\oplus} \mathbf{d}
$$

where $\mathbf{d}$ is a $n$-part composition obtained by the closure of the $\mathbb{R}^{n}$-vector which positive components are the $d_{k}$ 's. Note that a composition with equal components is the neutral element in $\mathbb{S}^{n}$ and this is the reason for cancelling the array containing the terms $\partial_{t} \rho$ in Eq. (7). Vectors in the right-hand side of Eq. (6) need additional definitions and properties. In the standard vector field analysis, the (spatial) gradient and divergence are useful differential linear operators. Previous Eq. (5-6) suggest that similar concepts can be defined for space-time simplicial fields. Definitions and some properties of such operators for $d=2$ follow.

Definition 3.1 (Simplicial (spatial) Gradient) Let $\mathbf{z}: S \times T \subseteq \mathbb{R}^{2} \times \mathbb{R} \rightarrow \mathbb{S}^{n}$ be a STSF. The simplicial (spatial) gradient is defined as a bivariate STSF, taking values in $\mathbb{S}^{n} \times \mathbb{S}^{n}$, given by

$$
\operatorname{grad}^{\oplus} \mathbf{z}(s, t)=\left(\partial_{x}^{\oplus} \mathbf{z}(s, t), \partial_{y}^{\oplus} \mathbf{z}(s, t)\right)
$$

Directional derivatives can be expressed as a kind of $\mathbb{R}^{d}$-inner product of the simplicial gradient and the direction in which the directional derivative is taken. However, the fact that simplicial derivatives are in $\mathbb{S}^{n}$ and spatial directions are in $\mathbb{R}^{d}$, introduces notational intricacies.

Proposition 3.1 (Simplicial gradient and directional derivatives) Let $\mathbf{z}: S \times T \subseteq \mathbb{R}^{2} \times$ $\mathbb{R} \rightarrow \mathbb{S}^{n}$ be a STSF and $\vec{u}=\left(u_{x}, u_{y}\right)$ be a vector. Directional derivative and gradient satisfies

$$
\partial_{\vec{u}}^{\oplus} \mathbf{z}(s, t)=\vec{u} \odot \operatorname{grad}^{\oplus} \mathbf{z}(s, t)
$$

where $\odot$ is interpreted as a perturbation-linear combination (Egozcue et al. (2011))

$$
\vec{u} \odot \operatorname{grad}^{\oplus} \mathbf{z}(s, t)=u_{x} \odot \partial_{x}^{\oplus} \mathbf{z}(s, t) \oplus u_{y} \odot \partial_{y}^{\oplus} \mathbf{z}(s, t)
$$

and $\operatorname{grad}^{\oplus} \mathbf{z}$ has been decomposed in their two components $\partial_{x}^{\oplus} \mathbf{z}$ and $\partial_{y}^{\oplus} \mathbf{z}$.

Definition 3.2 (Simplicial derivative along a multiple vector field) Let $\mathbf{z}: S \times T \subseteq$ $\mathbb{R}^{2} \times \mathbb{R} \rightarrow \mathbb{S}^{n}$ be a STSF with positive components $z_{k}(k=1,2, \ldots, n)$. Let $\mathbf{v}=\left(\vec{v}_{1}, \vec{v}_{2}, \ldots, \vec{v}_{n}\right)$ be a multiple vector field, being $\vec{v}_{k}=\left(v_{k x}, v_{k y}\right), k=1,2, \ldots, n$. The simplicial derivative of $\mathbf{z}$ along the multiple vector field $\mathbf{v}$ is

$$
\partial_{\mathbf{v}}^{\oplus} \mathbf{z}(s, t)=\mathcal{C} \exp \left(\begin{array}{c}
\partial_{\vec{v}_{1}} \log \left(z_{1}(s, t)\right) \\
\partial_{\vec{v}_{2}} \log \left(z_{2}(s, t)\right) \\
\vdots \\
\partial_{\vec{v}_{n}} \log \left(z_{n}(s, t)\right)
\end{array}\right)
$$

where $\partial_{\vec{v}_{k}} \log \left(z_{k}(s, t)\right)=v_{k x} \partial_{x} \log z_{k}+v_{k y} \partial_{y} \log z_{k}$ is an inner product of $\vec{v}_{k}$ and $\operatorname{grad}\left(\log z_{k}\right)$ in $\mathbb{R}^{2}$.

The simplicial derivative along a multiple vector field is not linear in the simplex. A linear combination of compositions like $\left(\alpha_{1} \odot \mathbf{z}_{1}\right) \oplus\left(\alpha_{2} \odot \mathbf{z}_{2}\right)$ is not equal to the perturbation-linear combination of the two derivatives. However, it is linear in the simplex for linear combinations of multiple vector fields.

Definition 3.3 (Simplicial divergence) Let $\mathbf{z}_{1}, \mathbf{z}_{2}: S \times T \subseteq \mathbb{R}^{2} \times \mathbb{R} \rightarrow \mathbb{S}^{n}$ be two STSFs. The simplicial divergence of the pair $\left(\mathbf{z}_{1}, \mathbf{z}_{2}\right)$ is a composition in $\mathbb{S}^{n}$ given by

$$
\begin{aligned}
\operatorname{div}^{\oplus}\left(\mathbf{z}_{1}, \mathbf{z}_{2}\right) & =\partial_{x}^{\oplus} \mathbf{z}_{1} \oplus \partial_{y}^{\oplus} \mathbf{z}_{2} \\
& =\operatorname{clr}^{-1}\left[\partial_{x} \operatorname{clr}\left(\mathbf{z}_{1}\right)+\partial_{y} \operatorname{clr}\left(\mathbf{z}_{2}\right)\right]
\end{aligned}
$$

Taking into account previous definitions, and computing $\mathrm{clr}^{-1}$ of the vectors in the right-hand side of Eq. (6)

$$
\mathcal{C} \exp \left(\begin{array}{c}
-\partial_{\vec{v}_{1}} \log \left(\rho d_{1}\right)  \tag{8}\\
-\partial_{\vec{v}_{2}} \log \left(\rho d_{2}\right) \\
\vdots \\
-\partial_{\vec{v}_{n}} \log \left(\rho d_{n}\right)
\end{array}\right) \oplus \mathcal{C} \exp \left(\begin{array}{c}
-\operatorname{div}\left(\vec{v}_{1}\right) \\
-\operatorname{div}\left(\vec{v}_{2}\right) \\
\vdots \\
-\operatorname{div}\left(\vec{v}_{n}\right)
\end{array}\right)=\partial_{\mathbf{v}}^{\oplus} \mathbf{d} \oplus \operatorname{div}^{\oplus}\left(\mathbf{w}_{x}, \mathbf{w}_{y}\right)
$$

where $\mathbf{w}_{x}=\operatorname{clr}^{-1}\left(\mathbf{v}_{x}\right), \mathbf{w}_{y}=\operatorname{clr}^{-1}\left(\mathbf{v}_{y}\right)$; moreover, $\mathbf{v}_{x}, \mathbf{v}_{y}$ are $\mathbb{R}^{n}$-vectors grouping the first and second components of $\vec{v}_{k}$ for $k=1,2, \ldots, n$ respectively. Note that $\operatorname{clr}\left(\mathbf{w}_{x}\right)=\mathbf{v}_{x}-\mathbf{v}_{x 0}$, where $\mathbf{v}_{x 0}$ is a constant-component vector which components are the aritmetic mean of the components of $\mathbf{v}_{x}$; and similarly for $\operatorname{clr}\left(\mathbf{w}_{y}\right)$.

Hence, considering Eqs. (7) and (8), the simplicial mass continuity equation can be written as

$$
\begin{equation*}
\partial_{t}^{\oplus} \mathbf{d}=\partial_{\mathbf{v}}^{\oplus} \mathbf{d} \oplus \operatorname{div}^{\oplus}\left(\mathbf{w}_{x}, \mathbf{w}_{y}\right) . \tag{9}
\end{equation*}
$$

An important feature is that the overall density $\rho$ does not appear in Eq.(9), therefore, this equation is purely compositional. Note that, in general, this is not a linear equation in the simplex due to the non-linearity of $\partial_{\mathbf{v}}^{\oplus} \mathbf{d}$ with respect to $\mathbf{d}$.

Furthermore, if each fluid species is incompressible, that is $\rho_{k}$ is constant, then Eq. (9) reduces to

$$
\operatorname{div}^{\oplus}\left(\mathbf{w}_{x}, \mathbf{w}_{y}\right)=\mathbf{n} \quad, \quad \mathbf{n}=\mathcal{C}(1,1, \ldots, 1),
$$

quite similar to standard continuity equation for an incompresible fluid flow (White, 1991).
In some cases the simplicial fields $\mathbf{z}_{1}$ and $\mathbf{z}_{2}$ can be the two components of a simplicial gradient, that is, there is a simplicial field $\mathbf{w}$ such that $\left(\mathbf{z}_{1}, \mathbf{z}_{2}\right)=\operatorname{grad}^{\oplus} \mathbf{w}=\left(\partial_{x}^{\oplus} \mathbf{w}, \partial_{y}^{\oplus} \mathbf{w}\right)$. In these cases, the bivariate simplicial field $\left(\mathbf{z}_{1}, \mathbf{z}_{2}\right)$ is said to derive from the potential composition $\mathbf{w}$, again following the ideas of the standard vector analysis. An important differential operator in this situation is the Laplacian $\Delta=\partial_{x}^{2}+\partial_{y}^{2}$. The compositional counterpart of the Laplacian can be defined as follows.

Definition 3.4 (Simplicial Laplacian) Let $\vec{x}$ be a location in $R$ and $\mathbf{w}$ be a STSF defined in a neighbourhood of $(\vec{x}, t)=(x, y, t)$. The simplicial Laplacian of $\mathbf{w}$ is a composition in $\mathbb{S}^{n}$ given by

$$
\Delta^{\oplus} \mathbf{w}=\operatorname{div}^{\oplus}\left(\operatorname{grad}^{\oplus} \mathbf{w}\right)=\partial_{x}^{\oplus 2} \mathbf{w} \oplus \partial_{y}^{\oplus 2} \mathbf{w}
$$

where the symbol $\partial^{\oplus 2}$ is the second order simplicial derivative (Egozcue et al., 2011).

To show the relevance of previous definitions, it is worth to state a compositional extension of the Gauss divergence theorem, here stated in $\mathbb{R}^{2}$.

THEOREM 3.1 Let $\mathbf{z}_{1}, \mathbf{z}_{2}: S \times T \subseteq \mathbb{R}^{2} \times \mathbb{R} \rightarrow \mathbb{S}^{n}$ be two STSF's, differentiable up to second order. Let $R$ be a bounded and connected domain in the plain with piecewise regular and closed boundary $\Gamma$. Consider $x, y$ as the Cartesian plain coordinates, and a piecewise regular parametrization $x=x(u)$, $y=y(u)$ of the boundary $\Gamma$. Then,

$$
\iint_{R}^{\oplus}(d x d y) \odot \operatorname{div}^{\oplus}\left(\mathbf{z}_{1}(x, y, t), \mathbf{z}_{2}(x, y, t)\right)=\int_{\Gamma}^{\oplus} d s \odot\left(\partial_{\vec{n}_{x}} \mathbf{z}_{1}(x(u), y(u), t) \oplus \partial_{\vec{n}_{y}} \mathbf{z}_{2}(x(u), y(u), t)\right)
$$

where $\vec{n}_{x}=-y^{\prime}(u) / \sqrt{x^{\prime}(u)^{2}+y^{\prime}(u)^{2}}, \vec{n}_{y}=x^{\prime}(u) / \sqrt{x^{\prime}(u)^{2}+y^{\prime}(u)^{2}}$ are, respectively, the vector fields of the first and second components of the normal direction to the boundary $\Gamma$; and ds $=$ $\sqrt{x^{\prime}(u)^{2}+y^{\prime}(u)^{2}} d u$.

## 4 Conclusions

The study of space-time simplicial fields reveals some interesting aspects and properties for applications in problems related with space-time evolutionary compositions. The present contribution is not complete and needs to be developed further.

It is possible to define, in a natural way, simplicial differential operators, similar to standard equations appearing in fluid mechanics and vector fields in general. The continuity equation of mass in fluid mechanics has been studied in some detail as a motivation to introduce some definitions. However, one of the needed simplicial differential operators, is not linear in the simplex, thus introducing features which are not dealt with in the standard formulation. It seems possible and useful to study the simplicial version of some important equations in Fluid Mechanics and other parts of Physics, such as advection-diffusion, Navier-Stokes and others in their compositional formulation.

## Acknowledgements

This research has been supported by the Spanish Ministry of Education and Science under project 'METRICS' (Ref. MTM2012-33236); and from the Agència de Gestió d’Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project Ref. 2009SGR424.

This research has been supported by the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) of the Generalitat de Catalunya (GENCAT) under the project "Compositional and Spatial Analysis" (COSDA, Ref: 2014SGR551; 2014-2016).

## References

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. London (UK): Chapman \& Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Egozcue, J. J., C. Barceló-Vidal, J. A. Martín-Fernández, E. Jarauta-Bragulat, J. L. DíazBarrero, and G. Mateu-Figueras (2011). Elements of simplicial linear algebra and geometry. See Pawlowsky-Glahn and Buccianti (2011).

Egozcue, J. J., E. Jarauta-Bragulat, and J. L. Díaz-Barrero (2011). Calculus of simplex-valued functions. See Pawlowsky-Glahn and Buccianti (2011).

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

Landau, L. D. and E. M. Lifshitz (1987). Fluid Mecahnics (2nd ed.), Volume 6 of Course of Theoretical Physics. Pergamon Press, New York. 531pp.

Pawlowsky-Glahn, V. and A. e. Buccianti (Eds.) (2011). Compositional Data Analysis: Theory to Applications. Wiley \& Sons.

White, F. M. (1991). Viscous fluid flow. Mechanical Engineering. McGraw Hill International Editions. 614 pp.

# Shifted-Dirichlet Regression vs Simplicial Regression: a comparison 

G.S. Monti ${ }^{1}$, G. Mateu-Figueras ${ }^{2}$, V. Pawlowsky-Glahn ${ }^{2}$, and J.J. Egozcue ${ }^{3}$<br>${ }^{1}$ Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy gianna.monti@unimib.it<br>${ }^{2}$ Department of Computer Science, Applied Mathematics, and Statistics, University of Girona, Spain<br>${ }^{3}$ Department of Applied Mathematics III, Technical University of Catalonia, Spain


#### Abstract

In Monti et al. (2014) a compositional model based on Scaled-Dirichlet regression was studied, following the approach suggested by Campbell and Mosimann (1987), and also studied by Hijazi and Jernigan (2009). The Scaled-Dirichlet distribution (Monti et al., 2011) is one of the generalizations of the Dirichlet distribution. Using an approach based on the Aitchison geometry and a suitable measure on the unit simplex $\mathcal{S}^{D}$, Monti et al. (2011) show that the Scaled-Dirichlet distribution can be viewed as the distribution of a perturbed random composition with Dirichlet density. The Scaled-Dirichlet regression model is obtained by allowing its parameters to change linearly with covariates, in order to assess the effects of covariates on the relative contributions of different components in a composition.


## 1 Introduction

Compositional data are vectors of parts of some whole which carry relative information. They are frequently represented as proportions or percentages, which are subject to a constant sum constraint, i.e. 1 or 100 , respectively. Their sample space is the unit simplex, denoted by

$$
\mathcal{S}^{D}=\left\{\mathbf{x}=\left(x_{1}, \ldots, x_{D}\right), x_{i}>0, \sum_{i=1}^{D} x_{i}=k\right\},
$$

The simplex $\left(\mathcal{S}^{D}, \oplus, \odot,\langle\cdot, \cdot\rangle_{a}\right)$ has a ( $D-1$ )-dimensional real Euclidean vector space structure, where $\oplus$ and $\odot$ are the pertubation and the powering operations and $\langle\cdot, \cdot\rangle_{a}$ represents the inner product. The geometry of the simplex is called simplicial or Aitchison geometry (Pawlowsky-Glahn and Egozcue, 2001).

There are many applied fields in which these kind of data are involved: from geology to election forecast, from medicine to economic studies.

Historically, there were different approaches to regression models which relate a compositional response variable with a system of covariates. The Dirichlet covariate model is one of them. It assumes that the response variable follows a Dirichlet distribution whose parameters are a loglinear function of some covariates (Campbell and Mosimann, 1987; Hijazi and Jernigan, 2009; Gueorguieva et al., 2008; Melo et al., 2009). As particular case, when the variable of interest is continuous and restricted to the unit interval $(0,1)$, i.e. when $D=2$, the Dirichlet regression is called Beta regression (Ferrari and Cribari-Neto, 2004). Other solutions involve log-ratio transformations of the response variable using normality assumptions (Aitchison and Shen, 1980; Aitchison, 1986; Egozcue et al., 2012), or models based on the generalized Liouville distribution (Rayens and Srinivasan, 1994).

The Scaled-Dirichlet covariate model is an extension of the Dirichlet one, based on the algebraic geometric structure of the simplex. The assumption is that $\mathbf{x}=\left(x_{1}, \ldots, x_{D}\right)$ is a compositional response vector, with $D$ components having Scaled-Dirichlet distribution $\mathcal{S D}^{\mathcal{D}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$, in which the parameters are allowed to change with a set of covariates.

The paper is structured as follows: section 2 gives a brief overview of the Scaled-Dirichlet distribution called in this paper Shifted-Dirichlet distribution. Section 3 defines the Shifted-Dirichlet
covariate model and deals with the issue of parameter estimation. An introduction to the alternative way based on simplicial regression is given in Section 4. Section 5 presents an example of application of the two techniques to real data.

## 2 Shifted-Dirichlet distribution

The Dirichlet distribution is one of the well known probability models suitable for random compositions. A random vector $\mathbf{X}=\left(X_{1}, \ldots, X_{D}\right) \in \mathcal{S}^{D}$ has a Dirichlet distribution, indicated by $\mathbf{X} \sim \operatorname{Dir}(\boldsymbol{\alpha})$, when its density function (with respect to the Aitchison measure) is:

$$
\begin{equation*}
f(\mathbf{x} ; \boldsymbol{\alpha})=\frac{\sqrt{D} \Gamma\left(\alpha_{+}\right)}{\prod_{i=1}^{D} \Gamma\left(\alpha_{i}\right)} \prod_{i=1}^{D} x_{i}^{\alpha_{i}} \tag{1}
\end{equation*}
$$

where $\alpha_{+}=\sum_{i=1}^{D} \alpha_{i}$, and $\Gamma$ denotes the gamma function. The Dirichlet distribution has $D$ parameters $\alpha_{i}$, which are assumed to be positive. One of its generalizations is the Scaled-Dirichlet distribution. A random vector $\mathbf{X} \in \mathcal{S}^{D}$ has a Scaled-Dirichlet distribution with parameters $\boldsymbol{\alpha}=$ $\left(\alpha_{1}, \ldots, \alpha_{D}\right) \in \mathbb{R}_{+}^{D}$ and $\boldsymbol{\beta}=\left(\beta_{1}, \ldots, \beta_{D}\right) \in \mathcal{S}^{D}$ if its density function is

$$
\begin{equation*}
f(\mathbf{x}, ; \boldsymbol{\alpha}, ; \boldsymbol{\beta})=\frac{\Gamma\left(\alpha_{+}\right) \sqrt{D}}{\prod_{i=1}^{D} \Gamma\left(\alpha_{i}\right)} \frac{\prod_{i=1}^{D}\left(\frac{x_{i}}{\beta_{i}}\right)^{\alpha_{i}}}{\left(\sum_{i=1}^{D} \frac{x_{i}}{\beta_{i}}\right)^{\alpha_{+}}} \tag{2}
\end{equation*}
$$

The density (2) is expressed with respect to the Aitchison probability measure (Pawlowsky-Glahn, 2003). See Monti et al. (2011) for a large discussion about the reasons and implications to use the Aitchison measure. This distribution will be denoted by $\mathbf{X} \sim \mathcal{S D}^{D}(\boldsymbol{\alpha}, \boldsymbol{\beta})$.

The number of parameters of this model is $2 D-1$. The Scaled-Dirichlet distribution can be obtained by normalising a vector of $D$ independent, scaled, gamma r.v.s $W_{i} \sim G a\left(\alpha_{i}, \beta_{i}\right), i=1,2, \ldots, D$; i.e. if $\mathbf{X}=\mathcal{C}(\mathbf{W})$, with $\mathbf{W} \in \mathbb{R}_{+}^{D}$, then $\mathbf{X} \sim \mathcal{S D}^{D}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ (Monti et al., 2011). This distribution can also be obtained as a perturbed random composition with a Dirichlet density. Let $\widetilde{\mathbf{X}} \sim \mathcal{D}^{D}(\boldsymbol{\alpha})$ be a random composition defined in $\mathcal{S}^{D}$, and let $\boldsymbol{\beta} \in \mathcal{S}^{D}$ be a composition. The random composition $\mathbf{X}=\ominus \boldsymbol{\beta} \oplus \widetilde{\mathbf{X}}$ has a $\mathcal{S D}^{D}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ distribution (note that $\ominus$ is the inverse operation of $\oplus$ ). Thus, taking into account the algebraic-geometric structure of the simplex, the Scaled-Dirichlet density is just a translation or shift of a Dirichlet density in the simplex. For this reason, it will be called hereafter Shifted-Dirichlet distribution. The expected value of $\mathbf{X} \sim \mathcal{S D}^{D}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ with respect to the Aitchison measure is

$$
\begin{equation*}
\mathrm{E}_{a}(\mathbf{X})=\ominus \boldsymbol{\beta} \oplus \mathrm{E}_{a}(\widetilde{\mathbf{X}}) \tag{3}
\end{equation*}
$$

where $\mathrm{E}_{a}(\widetilde{\mathbf{X}})$ is the expected value of a Dirichlet composition:

$$
\begin{equation*}
\mathrm{E}_{a}(\widetilde{\mathbf{X}})=\mathcal{C}\left(e^{\psi\left(\alpha_{1}\right)}, \ldots, e^{\psi\left(\alpha_{D}\right)}\right) \tag{4}
\end{equation*}
$$

where $\psi$ is the digamma function. The metric variance of $\mathbf{X}$ coincides with the metric variance of a Dirichlet composition, because this measure of dispersion is invariant under perturbation:

$$
\begin{equation*}
\operatorname{Mvar}(\mathbf{X})=\frac{D-1}{D}\left(\psi^{\prime}\left(\alpha_{1}\right), \ldots, \psi^{\prime}\left(\alpha_{D}\right)\right) \tag{5}
\end{equation*}
$$

where $\psi^{\prime}$ is the trigamma function.

## 3 Shifted-Dirichlet Regression

Linear regression with compositional response can be stated as follows. A compositional sample in $\mathcal{S}^{D}$, denoted by $\mathbf{x}_{1}, \ldots, \mathbf{x}_{n}$, is available. The sample size is $n$. Each data-point, $\mathbf{x}_{j},(j=$
$1, \ldots, n)$ is associated with one or more external variables or covariates grouped in the vector $\mathbf{s}_{j}=\left(s_{j 0}, s_{j 1}, \ldots, s_{j m}, \ldots, s_{j p}\right)$, where $s_{j 0}=1$ by convention. In this approach we hypothesize that each compositional observation $\mathbf{x}_{j}$ follows a conditional Shifted-Dirichlet distribution, given the set of covariates. Polynomial regression on a covariate $s$ is included as a particular case taking $s_{j m}=s_{j}^{m}$.
Following the approach of Campbell and Mosimann (1987) and Hijazi and Jernigan (2009), in order to incorporate in the model the covariate effects, we re-parametrise each parameter $\alpha_{i}$ of the density written in Equation (2) in terms of covariates and regression coefficients via the following log-linear model:

$$
\begin{equation*}
\alpha_{i j}=\alpha_{i j}\left(\mathbf{s}_{j}\right)=\exp \left\{\sum_{m=0}^{p} \delta_{i m} s_{j m}\right\} \tag{6}
\end{equation*}
$$

where $\mathbf{s}_{j}$ is the covariate vector recorded on the $j$-th observed composition $(j=1, \ldots, n)$, and $\delta_{i m}$ are the coefficients for the $m$-th covariate. The parameter $\delta_{i m}$ theoretically can vary by component, and the covariates may or may not be the same set of explanatory variables for each. We augment each vector $\mathbf{s}_{j}$ with 1 as first position for notation simplicity. Thus, given a sample of size $n, \mathbf{x}_{1}, \ldots, \mathbf{x}_{j}, \ldots, \mathbf{x}_{n}$, of identically and independently distributed compositional observations, the log-likelihood function for the re-parametrised Shifted-Dirichlet, given the covariates $\mathbf{s}$ and ignoring the part that does not involve the parameters, is proportional to ()

$$
\begin{align*}
l(\boldsymbol{\beta}, \boldsymbol{\delta} \mid \mathbf{x}, \mathbf{s}) \propto & \sum_{j=1}^{n} \log \Gamma\left(\sum_{i=1}^{D} \exp \left\{\sum_{m=0}^{p} \delta_{i m} s_{j m}\right\}\right)-\sum_{i=1}^{D} \log \Gamma\left(\exp \left\{\sum_{m=0}^{p} \delta_{i m} s_{j m}\right\}\right) \\
& -\sum_{i=1}^{D} \exp \left\{\sum_{m=0}^{p} \delta_{i m} s_{j m}\right\} \log \beta_{i}+\sum_{i=1}^{D} \exp \left\{\sum_{m=0}^{p} \delta_{i m} s_{j m}\right\} \log x_{i j}  \tag{7}\\
& -\left(\sum_{i=1}^{D} \exp \left\{\sum_{m=0}^{p} \delta_{i m} s_{j m}\right\}\right) \log \left(\sum_{i=1}^{D} \frac{x_{i}}{\beta_{i}}\right)
\end{align*}
$$

Equation (7) can be estimated using the maximum likelihood method via some optimisation algorithm, e.g. the Newton-Raphson algorithm. The choice of the starting value for the algorithm is of fundamental importance to get fast convergence. For the Dirichlet regression Hijazi and Jernigan (2009) proposed a method based on re-sampling from the original data; for each re-sample a Dirichlet model with constant parameters is fitted and the mean of the corresponding covariates is computed. After that, $D$ models of the form $\sum_{m=0}^{p} \delta_{i m} s_{j m}$ are fitted by least squares. The fitted coefficients $\hat{\delta}_{i m}$ are used as starting values.

Model selection can be performed by testing

$$
\begin{equation*}
H_{0}: \delta_{i m}=0 \tag{8}
\end{equation*}
$$

for some pair $(i, m), i=1, \ldots, D$ and $m=1, \ldots, p$. For it, the traditional likelihood ratio test needs to be implemented.

## 4 Simplicial regression

In order to explain the rationale of simplicial regression (Egozcue et al., 2012), it is necessary to recall the essential elements of the algebra for the simplex. For any vector of $D$ strictly positive real components,

$$
\mathbf{z}=\left(z_{1}, \ldots, z_{D}\right) \in \mathbb{R}_{+}^{D} \quad z_{i}>0, \text { for all } i=1, \ldots, D
$$

the closure operation of $\mathbf{z}$ is defined as

$$
\begin{equation*}
\mathcal{C}(\mathbf{z})=\left(\frac{k z_{1}}{\sum_{i=1}^{D} z_{i}}, \ldots, \frac{k z_{D}}{\sum_{i=1}^{D} z_{i}}\right) \in \mathcal{S}^{D} \tag{9}
\end{equation*}
$$

The two basic operations required for a vector space structure of the simplex are perturbation: given two compositions $\mathbf{x}$ and $\mathbf{y} \in \mathcal{S}^{D}$,

$$
\begin{equation*}
\mathbf{x} \oplus \mathbf{y}=\mathcal{C}\left(x_{1} y_{1}, \ldots, x_{D} y_{D}\right) \tag{10}
\end{equation*}
$$

and powering: given a composition $\mathbf{x} \in \mathcal{S}^{D}$ and a scalar $\alpha \in \mathbb{R}$,

$$
\begin{equation*}
\alpha \odot \mathbf{x}=\mathcal{C}\left(x_{1}^{\alpha}, \ldots, x_{D}^{\alpha}\right) . \tag{11}
\end{equation*}
$$

Furthermore, an inner product $\langle\cdot, \cdot\rangle_{a}$ is defined as

$$
\begin{equation*}
\langle\mathbf{x}, \mathbf{y}\rangle_{a}=\sum_{i=1}^{D} \ln \frac{x_{i}}{\mathrm{~g}_{\mathrm{m}}(\mathbf{x})} \ln \frac{y_{i}}{\mathrm{~g}_{\mathrm{m}}(\mathbf{y})}, \tag{12}
\end{equation*}
$$

where $\mathrm{g}_{\mathrm{m}}(\mathrm{x})$ denotes the geometric mean of the components of $\mathbf{x}$ (Billheimer et al., 2001; PawlowskyGlahn and Egozcue, 2001). This inner product induces a norm and a distance and a Euclidean vector space structure on the simplex.

Let $\left(\mathbf{e}_{1}, \ldots, \mathbf{e}_{D-1}\right)$ be an orthonormal basis of the simplex and consider the $(D-1) \times D$ matrix $\boldsymbol{\Psi}$ which rows are $\Psi_{i}=\operatorname{clr}\left(\mathbf{e}_{i}\right), i=1, \ldots, D-1$. This matrix is called contrast matrix associated with the orthonormal basis $\left(\mathbf{e}_{1}, \ldots, \mathbf{e}_{D-1}\right)$. Each row is called a $(\log )$ contrast.

The isometric logratio transformation, ilr for short, of $\mathbf{x}$ is the function ilr : $\mathcal{S}^{D} \rightarrow \mathbb{R}^{D-1}$, which assigns the coordinates $\mathbf{x}^{\star}$, with respect to the given basis, to the composition $\mathbf{x}$. The vector $\mathbf{x}^{\star}$ contains the $D-1$ ilr-coordinates of $\mathbf{x}$ in a Cartesian coordinate system. The inverse of the ilr-transformation is denoted ilr $^{-1}$. The function ilr is an isometry of vector spaces. The ilr transformation is computed as a simple matrix product: defining $\mathbf{V}=\boldsymbol{\Psi}^{\prime}$,

$$
\mathbf{x}^{\star}=\mathrm{ilr}(\mathbf{x})=\ln (\mathbf{x}) \mathbf{V}
$$

Inversion of ilr, i.e. recovering the composition from its coordinates is given by:

$$
\mathbf{x}=\operatorname{ilr}^{-1}\left(\mathbf{x}^{\star}\right)=\mathcal{C}\left(\exp \left(\mathbf{\Psi} \mathbf{x}^{\star}\right)\right)
$$

Given an orthonormal basis of the simplex, any composition $\mathbf{x} \in \mathcal{S}^{D}$ can be expressed as a linearcombination,

$$
\begin{aligned}
\mathbf{x} & =\left(x_{1}^{\star} \odot \mathbf{e}_{1}\right) \oplus\left(x_{2}^{\star} \odot \mathbf{e}_{2}\right) \oplus \ldots \oplus\left(x_{D-1}^{\star} \odot \mathbf{e}_{D-1}\right) \\
& =\bigoplus_{i=1}^{D-1}\left(x_{i}^{\star} \odot \mathbf{e}_{i}\right)
\end{aligned}
$$

where the symbol $\bigoplus$ represents repeated perturbation. The coefficients of the linear combination, for a fixed basis, are uniquely determined, given that in a Euclidean space any point can always be represented in a unique way by its coordinates with respect to an orthonormal basis. Once an orthonormal basis has been chosen, all standard statistical methods can be applied to coordinates and transferred to the simplex preserving their properties (Mateu-Figueras et al., 2011). This is the key point that is at the base of simplicial regression. The general model can be expressed as

$$
\begin{equation*}
\hat{\mathbf{x}}(\mathbf{s})=\left(s_{0} \odot \boldsymbol{\delta}_{0}\right) \oplus\left(s_{1} \odot \boldsymbol{\delta}_{1}\right) \oplus \ldots \oplus\left(s_{p} \odot \boldsymbol{\delta}_{p}\right)=\bigoplus_{m=1}^{p} s_{m} \odot \boldsymbol{\delta}_{m} \tag{13}
\end{equation*}
$$

The goal of estimating the coefficients $\boldsymbol{\delta}$ of a curve or surface in $\mathcal{S}^{D}$ is solved by translating it into a $D-1$ least squares problem, i.e. for each coordinate

$$
\begin{equation*}
\hat{\mathbf{x}}_{i}^{\star}(\mathbf{s})=\delta_{0 i}^{\star} s_{0}+\delta_{1 i}^{\star} s_{1}+\ldots+\delta_{p i}^{\star} s_{p} \quad i=1, \ldots, D-1, \tag{14}
\end{equation*}
$$

where $\boldsymbol{\delta}_{m}^{\star}=\left(\delta_{m 1}^{\star}, \ldots, \delta_{m, D-1}^{\star}\right)$ is the coordinate vector associated with $\boldsymbol{\delta}_{m}$. In the case of simple regression $m=1$ and $\mathbf{s}=s$, which is a straight-line in the simplex.

## 5 Example

In this section we compare the two different approaches to regression with a composition as dependent variable: the Shifted-Dirichlet covariate model and the simplicial regression. The example used is sediment data, i.e. (sand, silt, clay) compositions at different water depths at 39 locations within an Arctic lake. The goal is to model the effect of the depth covariate on the sediment composition. The data originally appeared in Coakley and Rust (1968), and were reproduced in Aitchison (1986).

To solve the maximisation problem defined in Equation (7) we used the BFGS algorithm implemented in R in the optim routine (R Core Team, 2014). To compute the parameters standard errors, the Hessian resulting from the optimisation process was used. As starting point, $\boldsymbol{\beta}=\mathcal{C}(1,1,1)$ was fixed, and for $\boldsymbol{\delta}$ the estimates were obtained fitting a simple Dirichlet regression:

$$
\ln \left(\alpha_{i}(s)\right)=\delta_{i 0}+\delta_{i 1} s \quad i=1,2,3
$$

The estimated parameters and their standard error are given in Table 1 and the fitted models (see Equation 3) are shown in Figure 1. Figures (1) and (2) shows the relationships between the

Table 1: Maximum likelihood estimates of Shifted-Dirichlet covariate model

| Parameter | Estimate | S.E. |
| :--- | :--- | :--- |
| $\delta_{10}$ | -0.601 | 0.369 |
| $\delta_{11}$ | -0.009 | 0.007 |
| $\delta_{20}$ | -1.005 | 0.365 |
| $\delta_{21}$ | 0.004 | 0.006 |
| $\delta_{30}$ | -1.474 | 0.341 |
| $\delta_{31}$ | 0.011 | 0.006 |
| $\beta_{1}$ | 0.257 | 0.232 |
| $\beta_{2}$ | 0.428 | 0.357 |



Figure 1: Estimated conditional means for the Shifted-Dirichlet covariate model of the three components for the Arctic Lake example.
compositions and the water depth estimated by Shifted-Dirichlet covariate model.


Figure 2: Evolution diagrams of the three proportions for the Arctic Lake example.

In order to apply the simplicial regression, ilr coordinates of the Arctic dataset were computed using a sequential binary partition (Egozcue and Pawlowsky-Glahn, 2005). Note that all chosen balances increase with depth (see Table (2)).

Table 2: Estimation coefficients for the ilr coordinates in simplicial regression for Arctic data

|  |  | sand-silt/clay | sand/silt |
| :---: | :--- | ---: | ---: |
| Intercept | Estimate | 1.87 | 0.459 |
|  | Std. Error | 0.263 | 0.167 |
| Slope | Estimate | -0.036 | -0.027 |
|  | Std. Error | 0.005 | 0.003 |

Predictions of the three coordinates can be back-transformed with the inverse ilr, to obtain a prediction of the proportions themselves (see Figure 2).

### 5.1 Model comparison

In order to assess the adequacy of the two regression approaches we examine the residuals behavior and some goodness-of-fit measures.

One suitable $R^{2}$ measure to evaluate the proportion of explained variation in the compositions by the covariate is based on total variability (Aitchison, 1986; Hijazi, 2006), based on the variation matrix of the transformed logratio data

$$
\begin{equation*}
\mathbf{T}(\mathbf{x})=\left[t_{i r}\right]=\left[\operatorname{var}\left(\ln \frac{x_{i}}{x_{r}}\right)\right] \quad i, r=1, \ldots, D \tag{15}
\end{equation*}
$$

Each element $t_{i r}$ is the usual variance of the logratio of parts $i$ and $r$. Aitchison's total variability measure totvar( $\mathbf{x}$ ), i.e. a measure of global dispersion of a compositional sample, is defined as

$$
\begin{equation*}
\operatorname{tot} \operatorname{var}(\mathbf{x})=\frac{1}{2 D} \sum_{i, r} \operatorname{var}\left(\ln \frac{x_{i}}{x_{r}}\right)=\frac{1}{2 D} \sum_{i, r} t_{i r} \tag{16}
\end{equation*}
$$

and $R_{T}^{2}$ as

$$
\begin{equation*}
R_{T}^{2}=\frac{\operatorname{tot} \operatorname{var}(\hat{\mathbf{x}})}{\operatorname{tot} \operatorname{var}(\mathbf{x})} \tag{17}
\end{equation*}
$$

which compares the total variability of the observed data and the fitted data.
Moreover, the Aitchison distance of any two compositions $\mathbf{x}$ and $\mathbf{y} \in \mathcal{S}^{D}$ is defined as

$$
\begin{equation*}
d_{a}(\mathbf{x}, \mathbf{y})=\sqrt{\frac{1}{2 D} \sum_{i=1}^{D} \sum_{r=1}^{D}\left(\ln \frac{x_{i}}{x_{r}}-\ln \frac{y_{i}}{y_{r}}\right)^{2}} \tag{18}
\end{equation*}
$$

If one considers the center of a compositional matrix given by

$$
\begin{equation*}
\mathbf{g}(\mathbf{x})=\mathcal{C}\left(\left(\prod_{j=1}^{n} x_{1 j}\right)^{1 / n}, \ldots,\left(\prod_{j=1}^{n} x_{D j}\right)^{1 / n}\right) \tag{19}
\end{equation*}
$$

i.e. the closure of the geometric mean of each sampled component, the compositional total sum of squares (CSST) and the compositional sum of squared residuals (CSSE) are then given by

$$
\begin{align*}
& C S S T=\sum_{j=1}^{n} d_{a}^{2}\left(\mathbf{x}_{j}, \mathbf{g}\right)  \tag{20}\\
& C S S E=\sum_{j=1}^{n} d_{a}^{2}\left(\mathbf{x}_{j}, \hat{\mathbf{x}}_{j}\right)
\end{align*}
$$

In this way another $R^{2}$-measure based on the compositional sum of squares (Hijazi, 2006), as in ordinary least-squares regression, is:

$$
\begin{equation*}
R_{A}^{2}=1-\frac{C S S E}{C S S T} \tag{21}
\end{equation*}
$$

Table 3: Goodness of fit measures for the two different regression models.

|  | $R_{T}^{2}$ | $R_{A}^{2}$ |
| :---: | ---: | ---: |
| Shifted Dirichlet regression | 0.644 | 0.936 |
| Simplicial regression | 0.626 | 0.948 |

The two measures of goodness of fit reported in Table (3) show good percentages of explained variation by the Shifted-Dirichlet model as well as by Simplicial Regression.

Diagnostic plots are given in Figures (3) and (4).
Further research is needed to investigate how the regression model developed in this paper can be useful in practice.

## Acknowledgement

Research partially financially supported by the Italian Ministry of University and Research, FAR (Fondi di Ateneo per la Ricerca) 2012, by the Ministerio de Economía y Competividad under the project METRICS Ref. MTM2012-33236; and by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project Ref: 2014SGR551.


Figure 3: Residual plots for the Arctic Lake example (Shifted-Dirichlet covariate model).


Figure 4: Residual plots for the Arctic Lake example (Simplicial regression).

## REFERENCES

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. London, Chapman \& Hall Ltd. (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Aitchison, J. and S. M. Shen (1980). Logistic-normal distributions. Some properties and uses. Biometrika 67(2), 261-272.

Billheimer, D., P. Guttorp, and W. Fagan (2001). Statistical interpretation of species composition. Journal of the American Statistical Association 96, 1205-1214.

Campbell, G. and J. Mosimann (1987). Multivariate methods for proportional shape. pp. 10-17. ASA Proceedings of the Section on Statistical Graphics.

Coakley, J. P. and B. R. Rust (1968). Sedimentation in an Arctic lake. Journal of Sedimentary Petrology 38, 1290-1300.

Egozcue, J. J., J. Daunis-i-Estadella, V. Pawlowsky-Glahn, K. Hron, and P. Filzmoser (2012). Simplicial regression. The normal model. J App Prob Stat 6(1-2), 87-108.

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

Ferrari, S. and F. Cribari-Neto (2004). Beta regression for modelling rates and proportions. J App Stat 31(7), 799-815.

Gueorguieva, R., R. Rosenheck, and D. Zelterman (2008). Dirichlet component regression and its applications to psychiatric data. Comput Stat Data Anal 52(12), 5344-5355.

Hijazi, R. H. (2006). Residuals and diagnostics in dirichlet regression. Technical report, Department of Statistics. United Arab Emirates University.

Hijazi, R. H. and R. Jernigan (2009). Modelling Compositional Data Using Dirichlet Regression Models. J App Prob Stat 4, 77-91.

Mateu-Figueras, G., V. Pawlowsky-Glahn, and J. J. Egozcue (2011). The principle of working on coordinates. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), Compositional Data Analysis: Theory and Applications. John Wiley \& Sons, Ltd.

Melo, T. F., K. L. Vasconcellos, and A. J. Lemonte (2009). Some restriction tests in a new class of regression models for proportions. Comput Stat Data Anal 53(12), 3972-3979.

Monti, G., G. Mateu-Figueras, V. Pawlowsky-Glahn, and J. Egozcue (2014). Scaled-Dirichlet covariate models for compositional data. Proceedings of 47 th Scientific Meeting of the Italian Statistical Society.

Monti, G. S., G. Mateu-Figueras, and V. Pawlowsky-Glahn (2011). Notes on the Scaled Dirichlet Distribution, pp. 128-138. John Wiley \& Sons, Ltd.

Pawlowsky-Glahn, V. (2003). Statistical modelling on coordinates. In S. Thió-Henestrosa and J. A. Martín-Fernández (Eds.), Compositional Data Analysis Workshop - CoDaWork'03, Proceedings. University of Girona, Girona (Spain), ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork03/. CD-ROM.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment 15, 384-398.

R Core Team (2014). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

Rayens, W. S. and C. Srinivasan (1994). Dependence properties of generalized liouville distributions on the simplex. J Am Stat Ass 89 (428), pp. 1465-1470.

# Bayesian estimation of the orthogonal decomposition of a contingency table 

M.I. Ortego ${ }^{1}$, J.J. Egozcue ${ }^{1}$<br>${ }^{1}$ Departament de Matemàtica Aplicada III. Universitat Politècnica de Catalunya. Barcelona, Spain;<br>ma.isabel.ortego@upc.edu


#### Abstract

Contingency tables can be parametrized by probabilities of each cell in a multinomial sampling. These probabilities constitute the joint probability function of the two or more discrete random categorical variables. These probability tables have been previously studied from a compositional point of view. The compositional approach to the problem ensures coherence when analysing contingency sub-tables. The main results are: (1) given a probability table, the closest independent probability table is the product of their geometric marginals; (2) the probability table can be orthogonally decomposed into an independent table and an interaction table; (3) the departure of independence can be measured using simplicial deviance, which is the Aitchison square norm of the interaction table. In previous works, the analysis has been performed from a frequentist point of view. This contribution is aimed at providing a Bayesian assessment of the decomposition. The resulting model is log-linear one, which parameters are the centered log-ratio transformations of the geometric marginals and the interaction table. Using a Dirichlet prior distribution of multinomial probabilities, the posterior distribution of multinomial probabilities is again a Dirichlet distribution. Simulation of this posterior allows to study the distribution of marginal and interaction parameters, checking the independence of the observed contingency table and cell interactions.


The results corresponding to a two-way contingency table example are presented.
Key words: Interaction; Independence; Simplicial deviance; Multinomial sampling; Aitchison geometry of the simplex; Orthogonal decomposition

## 1 Introduction

Contingency tables have been studied for a long time. There are many examples, dating from the beginning of the XX-th century which afforded elementary, but relevant, questions about such kind of data (e.g. Yule, 1912). Along the XX-th century many advances have been achieved. The introduction of log-linear models (Nelder, 1974) and generalized linear models (McCullagh and Nelder, 1983; Nelder and Wedderburn, 1972) were important milestones in the study of contingency tables. From the seventies up to now many extensions, improvement of methods and generalisations have been presented, for instance, see Everitt (1977), Darroch et al. (1980), Chambers and Welsh (1993), or Goodman (1996). However, challenges are still pendent for a straightforward solution, specially for the study of $n$-way contingency tables.

From the compositional point of view, the contingency tables have been studied only recently, with the early precedent of Kenett (1983). In workshop CoDaWork 2008 (Girona, Spain) Egozcue et al. (2008) introduced a perturbation-decomposition model for tables of multinomial parameters, thus opening new possibilities of analysis. This contribution was followed by other compositional attempts and applications (Gallo, 2011; Fačevicová and Hron, 2013). The approach proposed in Egozcue et al. (ToAp) is a kind of log-linear model but it has some differences with the standard ones. The main differences are the way in which marginals are found and the definition of interactions.

Here, the model based on the orthogonal decomposition of multinomial contingency tables is used to carry out a Bayesian estimation of both close independent multinomial parameters and the subsequent interactions. As the Bayesian approach is a likelihood based methodology, zeroes in the tables are handled without further problems.

In Section 2 the main features of the model based on orthogonal decomposition of contingency tables are recalled. Some definitions of the Bayesian framework are introduced in Section 3. Examples are presented in Section 4.

## 2 Orthogonal decomposition model

Two-way contingency tables coming from a multinomial sampling are considered. They are generated by a row-classification into $I$ classes, and a column-classification made of $J$ classes. The total number, $N$, of classified individuals is then distributed on the $I \times J$ cells of the contingency table (CT) according to the classification. The number of individuals pertaining to the $i j$-cell is denoted $n_{i j}$ for $i=1,2, \ldots, I$, and $j=1,2, \ldots, J$. The whole contingency table containing these counts is denoted $\mathbf{N}$. The table $\mathbf{N}$, as an array of counting random variables, is assumed to be multinomial distributed and its corresponding probability parameters denoted by $p_{i j} i=1,2, \ldots, I$, and $j=1,2, \ldots, J$. When arranged in a probability table ( PT ), these probability parameters are called probability table (PT). The sample space of a CT, like $\mathbf{N}$, is $I \times J$ the non-negative integers restricted to add to $N$ and they are not conceived as compositional data, even when the frequencies $\mathbf{N} / N$ are computed. In fact, they can contain zero-counts and only can correspond to fractions with $N$ as denominator. Alternatively, the probability parameters $\mathbf{P}$ are considered compositional. This can be summarized as (a) $\mathbf{P}$ is in $\mathcal{S}^{D}, D=I \cdot J$; (b) perturbation and powering, denoted $\oplus, \odot$, are vector space operations, and the dimension of $\mathcal{S}^{D}$ is $D-1$; (c) the centered log-ratio (clr) transformation is defined and inner product, norm and distances in $\mathcal{S}^{D}$ are the ordinary Euclidean inner product, norm and distance of the clr transformed PT's. As well-known for $\mathcal{S}^{D}$ (Pawlowsky-Glahn and Egozcue, 2001), the simplex endowed with $\oplus, \odot$, and the Aitchison inner product is a $D-1$-dimensional Euclidean space. More explicitly, consider two PT's, $\mathbf{P}$ and $\mathbf{Q}$ and a real number $\alpha$. The perturbation $\mathbf{W}=\mathbf{P} \oplus \mathbf{Q}$ is a PT with entries $w_{i j}=p_{i j} q_{i j} / \sum_{k m} p_{k m} q_{k m}$. The $\alpha$-powering $\mathbf{W}=\alpha \odot \mathbf{P}$ is a PT with entries $w_{i j}=p_{i j}^{\alpha} / \sum_{k m} p_{k m}^{\alpha}$. The clr of a PT is an
$(I \times J)$-array, $\mathbf{V}=c \operatorname{lr}(\mathbf{P})$ which entries are

$$
v_{i j}=\log \left(p_{i j}\right)-\frac{1}{D} \sum_{k=1}^{I} \sum_{m=1}^{J} \log p_{k m}
$$

The inverse clr-transformation is $\mathbf{P}=\mathcal{C} \exp (\mathbf{V})$, where $\exp$ operates componentwise. Denoting $\operatorname{clr}(\mathbf{P})=\mathbf{V}$ and $\operatorname{clr}(\mathbf{Q})=\mathbf{W}$, the Aitchison inner product, $\langle\cdot, \cdot\rangle_{a}$, and distance, $\mathrm{d}_{a}(\cdot, \cdot)$, of PT's is

$$
\langle\mathbf{P}, \mathbf{Q}\rangle_{a}=\langle\mathbf{V}, \mathbf{W}\rangle=\sum_{i=1}^{I} \sum_{j=1}^{J} v_{i j} w_{i j} \quad, \quad \mathrm{~d}_{a}^{2}(\mathbf{P}, \mathbf{Q})=\sum_{i=1}^{I} \sum_{j=1}^{J}\left(v_{i j}-w_{i j}\right)^{2}
$$

where $\langle\cdot, \cdot\rangle$ denotes the ordinary Euclidean inner product of arrays.
In these definitions, commonly used in compositional data analysis, there are, at least, two key points. The first one is the interpretability of the perturbation. Perturbation of PT's correspond to apply the Bayes formula to a PT, containing prior probabilities, using a likelihood arranged as a PT, up to the closure operation. The second point, is that the subcompositional coherence (PawlowskyGlahn et al., 2015; Egozcue, 2009), is guaranteed. In the case of PT's, subcompositional coherence assures that distance between two sub-tables have Aitchison distance smaller than or equal to the distance between the parent PT's.

The main result in Egozcue et al. (2008, ToAp) is that, independent PTs constitute a $(I-1)(J-1)$ dimensional linear subspace of $\mathcal{S}^{D}$. This means that any PT P can be projected orthogonally on it. The consequence is that $\mathbf{P}$ is decomposed in a unique way as

$$
\begin{equation*}
\mathbf{P}=\mathbf{P}_{i n d} \oplus \mathbf{P}_{i n t} \quad, \quad \mathbf{P}_{i n d} \perp \mathbf{P}_{i n t} \tag{1}
\end{equation*}
$$

where $\mathbf{P}_{\text {ind }}$ is the projection of $\mathbf{P}$ on the independent subspace, and $\mathbf{P}_{\text {int }}$ is in the orthogonal complement. The PT $\mathbf{P}_{\text {int }}$ is called interaction PT. The independent PT is on its turn decomposed into two new PT's, called marginal PT's, which have equal rows and equal columns respectively. The independent PT is then decomposed as

$$
\begin{equation*}
\mathbf{P}_{i n d}=\left(\mathbf{1}_{I} \mathbf{r}^{\top}\right) \oplus\left(\mathbf{c} \mathbf{1}_{J}^{\top}\right) \tag{2}
\end{equation*}
$$

where $\mathbf{r}, \mathbf{c}$ are compositions in $\mathcal{S}^{J}$ and $\mathcal{S}^{I}$ respectively, and they are treated as column vectors for matrix notation. The symbols $\mathbf{1}_{I}$ and $\mathbf{1}_{J}$ are column-vectors, with $I$ and $J$ components respectively, all of them equal to 1 .

Equations (1), (2) can be transformed by taking clr and these equations take the form

$$
\begin{equation*}
\operatorname{clr}(\mathbf{P})=\operatorname{clr}\left(\mathbf{P}_{i n d}\right)+\operatorname{clr}\left(\mathbf{P}_{i n t}\right)=\mathbf{1}_{I}(\operatorname{clr}(\mathbf{r}))^{\top}+\operatorname{clr}(\mathbf{c}) \mathbf{1}_{J}^{\top}+\operatorname{clr}\left(\mathbf{P}_{i n t}\right) . \tag{3}
\end{equation*}
$$

It should be remarked that $\operatorname{clr}(\mathbf{r})$ and $\operatorname{clr}(\mathbf{c})$ are clr transformations of compositions in $\mathcal{S}^{J}$ and $\mathcal{S}^{I}$ respectively and they are not PT's.

The marginal row and column, $\mathbf{r}$ and $\mathbf{c}$ respectively, are obtained from $\mathbf{P}$ as the closed geometric means by columns and rows of $\mathbf{P}$ respectively. This feature indicates that the nearest independent PT , in the sense of Aitchison geometry in $\mathcal{S}^{D}$, is not obtained from the traditional (arithmetic) marginals. This is an important difference from common analysis of contingency tables. As a consequence, $\operatorname{clr}\left(\mathbf{P}_{i n d}\right)$ has the property that its arithmetic and geometric marginals are equal up to a closure; and the geometric marginals of $\mathbf{P}_{\text {int }}$ are neutral in the simplex (i.e. all their elements are equal).
The decomposition in (3) implicitly defines a log-linear model which is revealed after taking clr ${ }^{-1}$ in (3). The log-linear model is then

$$
\begin{equation*}
\mathbf{P}=\mathcal{C} \exp \left[\operatorname{clr}\left(\mathbf{P}_{i n d}\right)+\operatorname{clr}\left(\mathbf{P}_{i n t}\right)\right]=\mathcal{C} \exp \left[\mathbf{1}_{I}(\operatorname{clr}(\mathbf{r}))^{\top}+\operatorname{clr}(\mathbf{c}) \mathbf{1}_{J}^{\top}+\operatorname{clr}\left(\mathbf{P}_{i n t}\right)\right] \tag{4}
\end{equation*}
$$

where the parameters are the $J$-coefficients in $\operatorname{clr}(\mathbf{r})$, the $I$ coefficients in $\operatorname{clr}(\mathbf{c})$ and the $D=I \cdot J$ coefficients in $\operatorname{clr}\left(\mathbf{P}_{i n t}\right)$. However, coefficients of any clr add to zero, and the number of free parameters is $(J-1)+(I-1)+(I J-1)=I J+I+J-3$. The number of clr-parameters in (4) can be reduced to $I J+I+J-3$ using ilr-coordinates, but this strategy is not used here as the clr-parameters can be interpreted directly.

In order to interpret the results when the log-linear model is fitted to a CT, some derived parameters may be useful. When the norm $\left\|\mathbf{P}_{\text {int }}\right\|_{a}$ is null, $\mathbf{P}_{\text {int }}$ is the neutral element in $\mathcal{S}^{D}$ and $\mathbf{P}$ is an independent PT. Therefore, $\left\|\mathbf{P}_{i n t}\right\|_{a}^{2}$ is an overall measure of dependence which was named simplicial deviance. When considered relative to the Aitchison square norm of $\mathbf{P}$, it can be called relative simplicial deviance. The corresponding definitions are

$$
\begin{equation*}
\Delta^{2}(\mathbf{P})=\left\|\mathbf{P}_{i n t}\right\|_{a}^{2} \quad, \quad R_{\Delta}^{2}(\mathbf{P})=\frac{\left\|\mathbf{P}_{i n t}\right\|_{a}^{2}}{\left\|\mathbf{P}_{i n d}\right\|_{a}^{2}+\left\|\mathbf{P}_{i n t}\right\|_{a}^{2}} \tag{5}
\end{equation*}
$$

where $\left\|\mathbf{P}_{\text {ind }}\right\|_{a}^{2}+\left\|\mathbf{P}_{\text {int }}\right\|_{a}^{2}=\|\mathbf{P}\|_{a}^{2}$ due to the orthogonal decomposition (1). Remarkably, $\Delta^{2}(\mathbf{P})$ does not depend on the marginals of $\mathbf{P}$; such a property is not shared by $R_{\Delta}^{2}(\mathbf{P})$. However $R_{\Delta}^{2}(\mathbf{P})$ has clear interpretation based on the facts of $0 \leq R_{\Delta}^{2}(\mathbf{P}) \leq 1, R_{\Delta}^{2}(\mathbf{P})=0$ implies independence of $\mathbf{P}$, whereas $R_{\Delta}^{2}(\mathbf{P})=1$ indicates that the nearest independent PT to $\mathbf{P}$ is the neutral (uniform) PT, and it can be considered as a pure interaction PT.

In order to interpret the coefficients of $\mathbf{V}=\operatorname{clr}\left(\mathbf{P}_{\text {int }}\right)$ it should be taken into account that the simplicial deviance is decomposed

$$
\begin{equation*}
\Delta^{2}(\mathbf{P})=\left\|\mathbf{P}_{\text {int }}\right\|_{a}^{2}=\sum_{i=1}^{I} \sum_{j=1}^{J} v_{i j}^{2} \tag{6}
\end{equation*}
$$

so that each cell contributes to the simplicial deviance with $v_{i j}^{2}$ thus deserving the name of cell interaction. A way of presenting these cell interactions is computing their relative value to the simplicial deviance or expressing them as percent of contribution. However, the signs of $v_{i j}$ are important as they indicate whether the probability in the cell $p_{i j}$ is smaller than the probability predicted probability using $\mathbf{P}_{\text {ind }}$ (negative $v_{i j}$ ) or it is larger than this predicted probability (positive $v_{i j}$ ). It has been proposed to use an interaction array reporting in each cell the value $\operatorname{sign}\left(v_{i j}\right)\left(v_{i j}^{2} / \Delta^{2}(\mathbf{P})\right.$. Unfortunately, the values of $v_{i j}$ cannot be interpreted separately as they add to zero. The analyst should look for large absolute values in the interaction array coupled by positive-negative interactions. Cells interactions are then interpreted jointly as the origins of the interactions are frequently coupled.

## 3 Bayesian analysis

Assume that an $I \times J$ contingency table $\mathbf{N}$ has been observed as the result of a multinomial sampling. After adopting the log-linear model (4), the multinomial probabilities $p_{i j}$ can be expressed as functions of the clr's of the geometric marginals $\operatorname{clr}(\mathbf{r})=\mathbf{z}^{(r)}=\left(z_{1}^{(r)}, z_{2}^{(r)}, \ldots, z_{J}^{(r)}\right)$, $\operatorname{clr}(\mathbf{c})=\mathbf{z}^{(c)}=\left(z_{1}^{(c)}, z_{2}^{(c)}, \ldots, z_{I}^{(c)}\right)$, and the entries of $\mathbf{V}=\operatorname{clr}\left(\mathbf{P}_{i n t}\right)$ denoted $v_{i j}$. Hence, the likelihood of these parameters, given the observation has the form

$$
L\left(\mathbf{z}^{(r)}, \mathbf{z}^{(c)}, \mathbf{V} \mid \mathbf{N}\right)=K \cdot \prod_{i=1}^{I} \prod_{j=1}^{J} p_{i j}^{n_{i j}}
$$

where all $p_{i j}$ are functions of $\mathbf{z}^{(r)}, \mathbf{z}^{(c)}, \mathbf{V}$. In order to simplify the estimation procedure, a Dirichlet distribution (e.g. Aitchison, 1986) can be chosen as initial joint distribution of the $p_{i j}$. If the chosen parameters of the Dirichlet distribution are $a_{i j}>0$, the final or posterior distribution
of the parameters is again a Dirichlet distribution with parameters $p_{i j}+a_{i j}$ and, therefore, the posterior distribution is

$$
\begin{equation*}
f\left(\mathbf{z}^{(r)}, \mathbf{z}^{(c)}, \mathbf{V} \mid \mathbf{N}\right)=\frac{\Gamma\left(\sum_{k} \sum_{m} a_{k m}\right)}{\prod_{k} \prod_{m} \Gamma\left(a_{k m}\right)} \prod_{i=1}^{I} \prod_{j=1}^{J} p_{i j}^{n_{i j}+a_{i j}} \quad, \quad \sum_{k} \sum_{m} p_{i j}=1 \tag{7}
\end{equation*}
$$

The goals of the Bayesian procedure are, at least, three: (a) estimation of posterior distribution of parameters $\mathbf{z}^{(r)}, \mathbf{z}^{(c)}, \mathbf{V}$ and their marginal distributions; (b) checking the hypothesis of independence of the observed CT; (c) study the distribution of the cell interactions $v_{i j}$ and checking whether they can be considered null or not. These three tasks are hardly carried out using the explicit distribution (7). A way out consists of drawing independent realisations from (7), and then, studying the simulated sample of parameters thus accomplishing goal (a).

Checking independence of the observed CT is performed through a predictive $p$-value (Bayarri and Berger, 2000; Meng, 1994) as proposed in goal (a). Assume that for each possible set of posterior parameters, $\mathbf{z}_{0}^{(r)}, \mathbf{z}_{0}^{(c)}, \mathbf{V}_{0}$, a likelihood ratio test is carried out on the hypothesis

$$
\begin{equation*}
H_{0}: \mathbf{z}^{(r)}=\mathbf{z}_{0}^{(r)}, \mathbf{z}^{(c)}=\mathbf{z}_{0}^{(c)}, \mathbf{V}=\mathbf{0} \tag{8}
\end{equation*}
$$

using the statistics

$$
\begin{equation*}
\Lambda=-2 \log \left(\frac{L\left(\mathbf{z}_{0}^{(r)}, \mathbf{z}_{0}^{(c)}, \mathbf{V}=\mathbf{0} \mid \mathbf{N}\right)}{L\left(\hat{\mathbf{z}}^{(r)}, \hat{\mathbf{z}}^{(c)}, \hat{\mathbf{V}} \mid \mathbf{N}\right)}\right) \tag{9}
\end{equation*}
$$

where $\hat{\mathbf{z}}^{(r)}, \hat{\mathbf{z}}^{(c)}, \hat{\mathbf{V}}$ denote the maximum likelihood estimators based on the sample CT. Asymptotically with $N$, the statistics $\Lambda$ has distribution $\chi^{2}$ with degrees of freedom $I J+I+J-3$. This corresponds to the number of estimated parameters, compared with no free parameter in $H_{0}$. For each set of values $\mathbf{z}_{0}^{(r)}, \mathbf{z}_{0}^{(c)}, \mathbf{V}_{0}$, one $p$-value $\alpha_{p 0}$ is obtained. The $\alpha_{p 0}$, as a function of the observed CT, has uniform distribution under asymptotic conditions (Robins et al., 2000). A predictive $p$-value, $\alpha$, with asymptotic uniform distribution, is obtained using

$$
\begin{equation*}
\alpha=\Phi\left(\frac{1}{m} \sum_{i=1}^{m} \Phi^{-1}\left(\alpha_{p 0}\right)\right) \tag{10}
\end{equation*}
$$

where the sum goes through the $m$-simulated sample of parameters $\mathbf{z}_{0}^{(r)}, \mathbf{z}_{0}^{(c)}$; and $\Phi$ denotes the standard normal distribution function (Ortego, 2015). Small values of $\alpha$ suggest rejection of the independence $H_{0}$.

The assessment of the hypothesis that a single cell interaction $v_{i j}$ is null is performed using Bayesian discrepancy $p$-values (Gelman et al., 1996), that is, computing the posterior probability of $v_{i j} \leq 0$ across the simulated sample. When this $p$-value is small (near to zero), or large (near to 1 ), rejection of $v_{i j}$ is suggested. This accomplishes goal (c).

## 4 Example of a Bayesian estimation of the orthogonal decomposition of a contingency table

### 4.1 Marks in a subject

The marks obtained by $N=84$ students in a exam of a college-level statistics subject are considered. Theoretical and practical (mostly problems) questions in the exams are marked separatedly. In this context, we want to know if the performance in theoretical questions can be considered independent from the performance in practical questions.

Table 1: Two-way contingency table containing the marks of the may examination of 84 students. Mark of the theoretical part of the exam (rows) vs. mark of the practical part (columns).

| mark (theory)/mark (prob) | $\mathrm{A}[8.5,10]$ | $\mathrm{B}[7,8.5)$ | $\mathrm{C}[5,7)$ | $\mathrm{D}[0,5)$ |
| :---: | :---: | :---: | :---: | :---: |
| $\mathrm{A}[8.5,10]$ | 1 | 0 | 4 | 4 |
| $\mathrm{~B}[7,8.5)$ | 2 | 4 | 6 | 13 |
| $\mathrm{C}[5,7)$ | 0 | 3 | 11 | 25 |
| $\mathrm{D}[0,5)$ | 1 | 1 | 5 | 24 |

The results of the exam may be classified into four groups: $A, B, C, D$, corresponding to the numeric interval of Spanish marks over 10 points. The results corresponding to this group of students have been organized in a two-way table (Table 1). We assume that these marks have been observed as a result of a multinomial sampling with probabilities $p_{i j}$. A Bayesian framewok is chosen for the estimation of the table parameters $p_{i j}$. For simplicity, a joint Dirichlet distribution has been assumed for these probabilities.

A Dirichlet priori has been set for the multinomial probabilities. Then, the posterior distribution of these parameters corresponds again to a Dirichlet distribution (Eq. 7). A (large) sample of the posterior distribution has been obtained. This sample is used to describe the uncertainty of the parameter estimates and other quantities of interest derived from them. For this dataset, a sample of the posterior of length 10000 has been obtained (e.g. Table 2).

Table 2: Example of table of the posterior sample

|  | 1 | 2 | 3 | 4 |
| ---: | ---: | ---: | ---: | ---: |
| 1 | 0.01 | 0.00 | 0.06 | 0.08 |
| 2 | 0.01 | 0.03 | 0.04 | 0.17 |
| 3 | 0.00 | 0.04 | 0.10 | 0.24 |
| 4 | 0.02 | 0.02 | 0.02 | 0.16 |

The tables of the sample of the posterior distribution of probabilities should be properly treated, due to their compositional character. The clr coordinates of the cells for each table have been computed (e.g. Table 3). The row and column geometric marginals of the clr coordinates have also been obtained for each of the tables of the posteriori sample. Also, each of these tables has been decomposed in its independent (e.g. Table 4) and interaction table (e.g. Table 5) following Eq. 1. That is, a sample of independent and interaction tables has been obtained from the sample of posterior tables. This allows to describe the uncertainty of quantities of interest derived from them, such as deviance, $\Delta^{2}(\mathbf{P})$, relative deviance, $R_{\Delta}^{2}(\mathbf{P})$, among others.

Table 3: Clr-coordinates of an element of the posterior sample. Row and column geometric marginals.

|  | 1 | 2 | 3 | 4 | rmarg |
| ---: | ---: | ---: | ---: | ---: | ---: |
| 1 | -1.36 | -5.39 | 1.02 | 1.30 | -1.105 |
| 2 | -0.57 | 0.35 | 0.68 | 2.06 | 0.631 |
| 3 | -4.19 | 0.52 | 1.48 | 2.38 | 0.048 |
| 4 | -0.26 | 0.05 | -0.08 | 2.00 | 0.426 |
| cmarg | -1.594 | -1.117 | 0.776 | 1.936 |  |

The departure from independence for the two sets of marks of interest may be measured observing the simplicial deviance (square of the norm) of the interaction components of the clr-coordinates of the posterior tables (Eq. 5). Fig. 1 shows the histogram of this deviance corresponding to the obtained sample of interaction tables. As the deviance is a measure of dependence, $0 \leq \Delta^{2}$ $(\mathbf{P})<+\infty$, a visual comparison with the zero value (red line) is included. For the example marks,

Table 4: Independent component of the decomposition of clr-coordinates of an element of the posterior sample

|  | 1 | 2 | 3 | 4 |
| ---: | ---: | ---: | ---: | ---: |
| 1 | -2.70 | -2.22 | -0.33 | 0.83 |
| 2 | -0.96 | -0.49 | 1.41 | 2.57 |
| 3 | -1.55 | -1.07 | 0.82 | 1.98 |
| 4 | -1.17 | -0.69 | 1.20 | 2.36 |

Table 5: Interaction component of the decomposition of clr-coordinates of an element of the posterior sample

|  | 1 | 2 | 3 | 4 |
| :--- | ---: | ---: | ---: | ---: |
| 1 | 1.34 | -3.17 | 1.35 | 0.47 |
| 2 | 0.39 | 0.83 | -0.72 | -0.50 |
| 3 | -2.64 | 1.59 | 0.65 | 0.40 |
| 4 | 0.90 | 0.74 | -1.28 | -0.36 |

although the median value is low, the amount of variability in the deviance values points to lack of independence between the theoretical and practical marks.


Figure 1: Simplicial deviance of the clr-interaction decomposition . Red line: null value. Blue line: median.
The relative simplicial deviance $R_{\Delta}^{2}(\mathbf{P})$ may seem easier to interpret than the deviance, as $0 \leq$ $R_{\Delta}^{2}(\mathbf{P}) \leq 1$, but this interpretation should be taken with care, as this coefficient is not marginal invariant. Fig. 2 shows the histogram corresponding to the simplicial relative deviance of the interaction component of the clr-coordinates of the posterior sample with respect to the norm of the whole table. A zero-line is also included for a visual comparison. In this case, the majority of the relative deviance values are around 0.3 , being near to its median value, reassuring the interpretation of lack of independence.

The simplicial deviance is an overall measure of dependence, but often more detail is needed. The cell values of the interaction table (e.g. Table 6) provide this detail, but the direct interpretation may be incorrect, due to its compositional character. In order to obtain a detailed description of interaction using the apropiate scale, the cell interactions for the clr-coordinates of each table of the posterior sample have been computed. Fig. 3 shows the histogram of the cell interactions as a summary of the obtained results. For visual comparison, a zero line has been added to each histogram. Visually, a zero line near the middle of the histogram indicates no interaction added by that cell (e.g. histogram corresponding to cell 4). If the zero line is far from the center of the histogram (e.g. histogram corresponding to cell 1), that cell may be adding interaction to the


Figure 2: Simplicial Relative Deviance of the clr-interaction decomposition. Red line: null value. Blue line: median.
deviance (Eq. 6), and should be studied.


Figure 3: clr-cell interactions. Red lines: null interaction
However, for an easier understanding of the importance of each cell, the interaction array of the cells has been also computed (e.g. Table 6), measuring the percentage of interaction added to the deviance by each cell, and including the sign of this interaction. The histogram of the signed interaction array for the clr-coordinates of the posterior sample of tables is shown in Fig. 4. Visually, cells with interaction arrays clearly different from zero should be studied, as they are the influential ones. It seems that the most influential cells for the departure of independence are cells 1 and 3, namely 'A in theory' vs 'A in practical' marks and 'A in theory' vs ' C in practical' marks, with more or less the same weight and opposite signs (see Fig. 4, first row). The positive sign of the interaction array for cell 1 means that the predicted probability for the cell is larger than the predicted by the independent table, while the predicted probabilities for cell 3 (negative sign) are lower than the probabilities predicted by the independent table. Other cells, as cell 5 , are also influential, but with a lower weight. The hypothesis of null interaction has been also assessed by means of a Bayesian $p$-value based on a discrepancy (posterior probability of $v_{i j} \leq 0$ across the
sample) (Table 7). If the zero value is central in the sample, i.e. the proportion of $v_{i j} \leq 0$ is near 0.5 , the hypothesis is not rejected. Otherwise, small or large proportions, lead to the rejection of the null interaction hypothesis. For instance, for cell 3 the Bayesian $p$-value is 0.956 , and therefore the null hypothesis is clearly rejected. For cell 1 , the $p$-value is 0.101 and, although the value is low, the decision of rejection is not so straightforward.

Table 6: Example of interaction array

|  | 1 | 2 | 3 | 4 |
| :--- | ---: | ---: | ---: | ---: |
| 1 | 6.27 | 0.54 | -24.25 | 2.84 |
| 2 | -34.89 | 2.42 | 8.79 | 1.92 |
| 3 | 6.37 | -1.82 | 1.49 | -5.74 |
| 4 | 0.77 | -0.89 | 0.55 | -0.46 |



Figure 4: interaction array. Red line: null value. Blue line: median

Table 7: Bayesian $p$-value based on discrepancy (posterior probability of $v_{i j} \leq 0$ across the sample)

|  | 1 | 2 | 3 | 4 |
| ---: | ---: | ---: | ---: | ---: |
| 1 | 0.101 | 0.297 | 0.956 | 0.455 |
| 2 | 0.773 | 0.131 | 0.223 | 0.697 |
| 3 | 0.261 | 0.873 | 0.175 | 0.746 |
| 4 | 0.880 | 0.914 | 0.115 | 0.052 |

### 4.2 An independence test

Simplicial deviance, relative deviance and the interaction array are useful quantities to study dependence in a contingency table. However, it is usual to discuss independence in contingency tables by means of a test (e.g. Eq. 8). In our example, are the marks for theory and practice in the exam independent? In the established theoretical context, that can be rephrased as, does the contingency table of marks belong to the subspace of independent tables?

$$
H_{0}: T=\mathbf{P}_{i n d} \in \mathcal{S}_{i n d}^{D} \quad ; \quad H_{1}: \mathbf{P} \in \mathcal{S}_{i n d}^{D}
$$

The selected likelihood ratio test statistic (Eq. 9), is based on the sample of estimates of the independent component $\mathbf{P}_{i n d}, \widehat{\mathbf{P}}_{\text {ind }}$. For each table of the sample of the posterior distribution, its decomposition into independent and interaction component has been obtained in section 4.1. The proposed test statistic and its corresponding predictive $p$-value have been computed for each of these decompositions. This sample of $p$-values can be used to measure the uncertainty of the decision of the independence test. Fig. 5 shows the histogram of these predictive $p$-values for the posterior sample of tables. It can be observed that there is variability in the sample of $p$-values, with a majority of small values, leading to the rejection of the independence hypothesis. However, the lack of uniformity of $p$-values and their relative scale are problematic for their interpretation (Robins et al. (2000)). Therefore, the predictive $p$-values of the sample have been suitably transformed and combined, in order to obtain a summary $p$-value, $\alpha$, with asymptotic uniform distribution. In this case, $\alpha$ is nearly 0 , and the independence hypothesis has been rejected, as already pointed out by the deviance values. That is, it cannot be considered that the theory and practical marks of this exam are independent.


Figure 5: $p$-value of the multinomial likelihood ratio independence test

## 5 Conclusions

Contingency tables have been broadly studied, although only recently have been studied from the compositional point of view. The orthogonal decomposition of multinomial contingency tables has been presented. Also, a Bayesian framework for the estimation of the parameters of the tables has been presented, as a novelty in the compositional treatment of these tables.

A two-way contingency table containing marks from an exam of a college-level statistics course has been studied as an example. Results show that theory and practical exam marks cannot be considered as independent, as different characters of the table decomposition suggest. The Bayesian point of view allows to consider the uncertainty of the estimations of the different coefficients, very useful when counts in the original table are small. The fact that small and even zero counts are not problematic from this point of view, makes Bayesian estimation very useful in this context.

## Acknowledgements

This research has been partially funded by the Spanish Ministerio de Economía y Competitividad under project 'Metrics' Ref.MTM2012-33236; and by the Agència de Gestió d'Ajuts Universitaris
i de Recerca (AGAUR) of the Generalitat de Catalunya under the project "Compositional and Spatial Analysis" (COSDA) (Ref: 2014SGR551;2014-2016).

## REFERENCES

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman \& Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Bayarri, M. J. and J. O. Berger (2000). P-values for composite null models. Journal of the American Statistical Association 95, 1127-1142.

Chambers, R. L. and A. H. Welsh (1993). Log-linear models for survey data with non-ignorable non-response. J. R. Statist. Soc. B 55(1), 157-170.

Darroch, J. N., S. L. Lauritzen, and T. P. Speed (1980). Markov fields and log-linear interaction models for contingency tables. The Annals of Statistics 8(3), 522-539.

Daunis-i Estadella, J. and J. A. Martín-Fernández (Eds.) (2008). Proceedings of CODAWORK'08, The 3rd Compositional Data Analysis Workshop, Girona. University of Girona. ISBN 978-84-8458-272-4.

Egozcue, J. J. (2009). Reply to "On the Harker variation diagrams; ..." by J. A. Cortés. Mathematical Geosciences 41(7), 829-834.

Egozcue, J. J., J. L. Díaz-Barrero, and V. Pawlowsky-Glahn (2008). Compositional analysis of bivariate discrete probabilities. See Daunis-i Estadella and Martín-Fernández (2008), pp. 1-11. ISBN 978-84-8458-272-4.

Egozcue, J. J., V. Pawlowsky-Glahn, M. Templ, and K. Hron (ToAp). Independence in contingency tables using simplicial geometry. Communications in Statistics. Accepted. In press.

Egozcue, J. J., R. Tolosana-Delgado, and M. I. Ortego (Eds.) (2011). Proceedings of the 4th International Workshop on Compositional Data Analysis, Barcelona, Spain. Departament de Matemàtica Aplicada III (Universitat Politècnica de Catalunya): CIMNE. ISBN 978-84-87867-76-7.

Everitt, B. S. (1977). The Analysis of Contingency Tables. John Wiley \& Sons, Inc. New York, New York, USA.

Fačevicová, K. and K. Hron (2013, June 2013). Statistical analysis of compositional 2 X 2 tables. See Hron et al. (2013). ISBN 978-3-200-03103-6.

Gallo, M. (2011). Tree-way compositional data analysis. See Egozcue et al. (2011), pp. 1-13. ISBN 978-84-87867-76-7.

Gelman, A., X. L. Meng, and H. Stern (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). Statistica Sinica 6, 733-807.

Goodman, L. A. (1996). A single general method for the analysis of cross-classified data: Reconciliation and synthesis of some methods of Pearson, Yule, and Fisher, and also some methods of correspondence analysis and association analysis. Journal of the American Statistical Association 91 (433), 408-428.

Hron, K., P. Filzmoser, and M. Templ (Eds.) (2013, June 2013). Proceedings of the 5th International Workshop on Compositional Data Analysis. TU Wien: TU Wien. ISBN 978-3-200-03103-6.

Kenett, R. S. (1983). On an exploratory analysis of contingency tables. The Statistician 32(3), 395-403.

McCullagh, P. and J. A. Nelder (1983). Generalized Linear Models. Chapman and Hall, London, UK. 522 p.

Meng, X. L. (1994). Posterior predictive p-values. Annals of Statistics 22, 1142-1160.
Nelder, J. A. (1974). Log linear models for contingency tables: a generalization of classical least squares. Appl. Statist. 23, 323-329.

Nelder, J. A. and R. W. M. Wedderburn (1972). Generalized linear models. Journal of the Royal Statistical Society, series A 135, 370-384.

Ortego, M. I. (2015). Estimación Bayesiana de cópulas extremales en procesos de Poisson. Ph. D. thesis, Universitat Politècnica de Catalunya.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment (SERRA) 15(5), 384-398.

Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2015). Modeling and analysis of compositional data. Statistics in practice. John Wiley \& Sons, Chichester UK. 272 pp.

Robins, J. M., A. van der Vaart, and V. Ventura (2000). Asymptotic distribution of p-values in composite null models. Journal of the American Statistical Association 95(452), 1143-1156.

Yule, G. U. (1912). On the methods of measuring association between two attributes. Journal of the Royal Statistical Society 75, 579-642.

# Relationship between popularity of key words in the Google browser and the evolution of worldwide financial indexes 

R. Ortells ${ }^{1}$, J.J. Egozcue ${ }^{2}$, M.I. Ortego ${ }^{2}$, and A. Garola ${ }^{3}$<br>${ }^{1}$ Universitat Politècnica de Catalunya (UPC); robert.ortells@estudiant.upc.edu<br>${ }^{2}$ Applied Mathematics III Department, Universitat Politècnica de Catalunya (UPC);<br>${ }^{3}$ Infrastructure and Territory Department, Universitat Politècnica de Catalunya (UPC);

key words: financial markets, Google searches, stock market indexes, sovereign bond yields, commodities, compositional data, multiple linear regression.


#### Abstract

The purpose of this contribution is to evaluate whether there is enough statistical basis to establish a relationship between the popularity of certain terms in the Google, Inc. browser and the evolution of several worldwide economic indexes the subsequent week. A linear model trying to predict the evolution of 19 financial indexes from all over the world with the information of how many times a selected group of 200 key words are looked up online the previous week is proposed.

The linear model that is proposed takes a compositional approach due to two reasons. First, because the information contained in the values of the financial indexes has a compositional nature. The strongest proof supporting this idea is that in case all values for the indexes on a certain week were multiplied by a factor, the information would remain unchanged. In fact, the value for a certain index is irrelevant by itself, since it is its evolution with respect to the rest of indexes that indicates whether it is performing well. Therefore, this idea suggests that the numerical values of the 19 indexes for a certain week can be understood as a vector of the simplex and be analyzed accordingly.


Second, the explicative variable has to be understood as a vector of the simplex as well, for a similar reason as before. For instance, let us imagine that the number of times the words are looked up online in a certain week was multiplied by a factor. Indeed, the information contained in this vector would be exactly the same. Moreover, it seems intuitive as well how the absolute value for the number of searches is irrelevant by itself, since we will be interested in the relationships amongst variables.
For the reasons we have just set, a compositional approach seems necessary in order to address the problem successfully, since both the explicative and the predicted variable present a compositional nature. In other words, despite not adding up to a constant, the components of the vectors of both the explicative and predicted variable seem to be closely related in terms of giving information of a part of a whole, so tackling the problem through a compositional perspective seems appropriate.

The analysis consists of an exploratory analysis of both response (indexes) and explanatory (searches) variables and a compositional linear multiple regression between both sets of variables.

## 1 Introduction

The scope of the project is to analyze whether the weekly evolution of the popularity of a group of words in the Google browser is explicative of the behaviour of the financial markets on the following week. By the term financial markets, we understand a selection of economic indicators from all over the world, including the main stock market indexes, sovereign bond yields and commodity prices. This selection of indicators has been made assuming that the characterization of what we understand as global markets can be done through the most relevant financial centres of the world.

The potential usefulness of the searches in the Google browser (provided by Google, Inc. through the subsidiary Google Trends (Google Inc., 2014)) has been widely studied in the past. For example, Preis et al. (2013) tried to approach the correlation between the popularity of a group of selected words and the Dow Jones Industrial Average Index performance. Other economic-related lines of investigation have tried to address the correlation between the popularity of certain words in an Internet browser and other variables such as the stock market volume (Bordino et al., 2012; Choi and Varian, 2012). However, the information of the popularity of certain words on Google has also been used in multiple other fields apart from the economic world; for instance, Ginsberg et al. (2009) focused on the potential correlation of such popularity and the spreading of epidemics. According to the state of the art, the key message is that the information on the patterns of searches in the Google browser can be explicative of the behaviour of multiple indicators, as well as it can anticipate future trends. Indeed, according to the current lifestyle in most developed countries, the use of Internet seems to be extremely useful to understand not only how society behaves at present time but also to anticipate how it will behave in the future.

With reference to what the present document intends to add to the state of the art, we have tried to address a similar problem to the one considered in Preis et al. (2013) taking into account a compositional approach, both for the information of the popularity of the words as well as the economic indexes. First, a justification on why such an approach is appropriate has been made. After that, a regression model has been proposed. Finally, the results have been analyzed critically, in order to evaluate whether the compositional perspective has worked as expected.

## 2 Explicative and predicted variables

### 2.1 Selection and treatment of the 200 terms in the Google browser

The database containing the words that have been selected does not only include economic terms. Previous research such as Preis et al. (2013) has considered exclusively words from the economic world, but since a principal component analysis has been used, we have decided to include additional words to evaluate whether they can be explicative as well. These 200 words are presented in Table 1, and include the words that were used in Preis et al. (2013) plus an additional group that has been considered by the authors. It is important to point out that we have deliberately chosen universal terms. For example, we have avoided using words like subprime or Lehman Brothers because even though their popularity would have certainly been explicative of the behaviour of the markets during the period 2004-2014, they would have biased the results of our project.

Regarding the information provided by Google, Inc. (Google Inc., 2014), a few comments have to be made. First, the information on the number of searches on the browser is not the absolute number of searches of each word. For every word, the database associated to it has been normalized in such a way that the historical maximum appears as 100 , so there is no way to know the true absolute popularity of the words. Therefore, the information of the number of searches every week is expressed as a fraction of the historical maximum. Second, the information on the number of searches has been rounded to the nearest integer value. For example, in case after the normalization the value associated to one week is 62.3 , Google returns the rounded value 62 . Provided that no

Table 1: List of the 200 words that have been used in the study. Note that no capital letters have been used in the words except those representing acronyms formed with the first letter of each word. For example, the term "fed", commonly used to refer to the Federal Reserve, has been considered without capital letters, while the term "ECB" (which stands for European Central Bank) has been used in capital letters.

| debt | society | water | trader | interest rates | savings | shortage | free |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| color | leverage | rich | rare earths | business | speculation | capitalization | year |
| stocks | loss | risk | tourism | shares | credit default swap | utilities | university |
| restaurant | cash | gold | politics | mortgage | sovereign | transportation | pencil |
| portfolio | office | success | energy | S\&P500 | treasury | trading | book |
| inflation | fine | oil | consume | central bank | volatility | currency | travel |
| housing | stock market | war | consumption | futures | wage | high yield | dog |
| dow jones | banking | economy | freedom | commodities | salary | assets | disco |
| revenue | crisis | chance | dividend | brent | wealth | ECB | cinema |
| economics | happy | short sell | world | west texas | bank | yen | bus |
| credit | car | lifestyle | conflict | equity | exports | yuan | animal |
| markets | nasdaq | greed | kitchen | tax | imports | securities | safari |
| return | gains | food | forex | deflation | income | warrants | name |
| unemployment | finance | financial markets | home | GDP | capital | ETF | think |
| money | sell | movie | crash | IPO | hedge fund | CFD | boy |
| religion | investment | nyse | transaction | rating | loan | insolvency | old |
| cancer | fed | ore | garden | bund | eurozone | disaster | word |
| growth | house | opportunity | fond | bankruptcy | euro | fund | number |
| investment | metals | health | train | bullish | dollar | happiness | press |
| hedge | travel | short selling | labor | bearish | BRIC | sex | mother |
| marriage | returns | earnings | fun | productivity | terrorism | church | sun |
| bonds | gains | arts | environment | deficit | attack | shopping | eye |
| derivatives | default | culture | ring | devaluation | collapse | computer | life |
| headlines | present | bubble | recession | federal reserve | G20 | science | time |
| profit | holiday | buy | yield | liquidity | danger | film | person |

further information is provided by the browser, there is uncertainty on how the raw big data has been previously treated by Google, Inc.

The database containing the weekly amount of searches for the 200 words has been treated in order to, first, address the compositional nature of the data, and, second, to reduce the dimension according to a principal component analysis. First of all, the zeros on the database have been identified and addressed in the way that was proposed by Martín-Fernández et al. (2012) and Martín-Fernández et al. (2003). The decision of approaching the presence of zeros in this way comes from the strong belief that such zeros are rounding zeros (very small amount of searches compared to the historical maximum that appear as zeros even though they are not). After dealing with the zeros on the database, a compositional approach is considered. There are many explanations supporting the compositional nature of the database. First, in case the number of searches were all multiplied by a factor $k$, the information would remain unchanged. Second, the information for a certain word is irrelevant by itself since it is only explicative when compared to the rest. For instance, in case the word bankruptcy increases in popularity for a certain week and the other words present a similar increase, it seems that not much has happened. However, in case the searches increase for such word and remain quite the same for the rest, this may be relevant. For this reason, it seems appropriate to understand the number of searches for a certain word as a part of a whole and apply a compositional treatment accordingly. However, we should be aware that the sum of the searches for a certain week is not constant, and is unknown. Indeed, the total number of times that the 200 words are looked up is different every week, but this is irrelevant because we are not interested in working with the raw data. It is important to highlight that this detail makes the present study slightly different from the traditional compositional analysis, since even though it seems intuitive to understand the information as parts of a whole the actual whole is not only variable for each week but also unknown.

Once the compositional treatment has been justified, we have addressed the fact that the information on the number of searches has been normalized in such a way that the historical maximum has become 100. For a certain week $i$, we have a 200 component vector containing the information of the popularity of each word:

$$
X_{i}=\left[x_{1}, \ldots, x_{200}\right]
$$

Table 2: Cumulated explained variance by the first 20 principal components of the data set.

| Component | Cumulated variance | Component | Cumulated variance |
| :---: | :---: | :---: | :---: |
| Comp. 1 | $53.3 \%$ | Comp. 11 | $89.5 \%$ |
| Comp. 2 | $65.2 \%$ | Comp. 12 | $90.4 \%$ |
| Comp. 3 | $72.3 \%$ | Comp. 13 | $91.3 \%$ |
| Comp. 4 | $76.7 \%$ | Comp. 14 | $92.0 \%$ |
| Comp. 5 | $79.6 \%$ | Comp. 15 | $92.6 \%$ |
| Comp. 6 | $82.2 \%$ | Comp. 16 | $93.1 \%$ |
| Comp. 7 | $84.2 \%$ | Comp. 17 | $93.5 \%$ |
| Comp. 8 | $85.8 \%$ | Comp. 18 | $93.9 \%$ |
| Comp. 9 | $87.2 \%$ | Comp. 19 | $94.3 \%$ |
| Comp. 10 | $88.4 \%$ | Comp. 20 | $94.6 \%$ |

We have already proved how this vector belongs to the simplex space Aitchison et al. (2002) due to its compositional nature. However, we can see how each component of the vector represents the absolute number of searches multiplied by an unknown normalizing factor such that the historical maximum becomes 100. This operation is equivalent to a perturbation in the simplex for each one of the 200 components, so the relative information contained in the vector remains undamaged (for further information on how perturbation does not alter the information of the vector see Aitchison (1986), Aitchison (1982) and Aitchison et al. (2002)). It is important to point out that the value for this normalizing factor is irrelevant, since we are not interested in obtaining the absolute number of searches.

The compositional treatment has consisted in computing the CLR coordinates of the dataset, according to Aitchison et al. (2002). For a certain week $i$, the CLR coordinates have one dimension less than the vector of searches, according to the principles of the compositional analysis. Therefore, after the CLR transformation the vector for each week becomes a vector of 199 components. After that, we have applied a singular value decomposition and obtained the principal components of the space. The ultimate scope of performing such operation is to reduce the dimension of the vector so we can decrease the number of explicative variables in the regression. Regarding the regression, we will only consider the first 15 components of the vectors resulting from the principal component analysis. The percentage of the cumulated explained variance up to the 20 th component is detailed in Table 2.

To sum up, the database containing the information on the searches of the words has been treated as compositional (through a CLR transformation), and has been simplified according to a singular value decomposition criterion. The final result has been a new database, where each week is characterized through a vector of 15 components (since we have limited the number of explicative variables to the 15 first components of the vectors), which will be our explicative variables in the regression model we will propose. All computations have been made with R software ( R Development Core Team, 2008).

### 2.2 Selection and treatment of the indexes

The evolution of the financial markets has been modelled through studying 19 key indexes, which are representative of stock market indexes and sovereign bonds yields from all over the world, as well as the most important commodity markets. The information on the values of the indexes has obtained from Yahoo Finance and Investing.com (Yahoo (2013) and Fusion Media Limited (2014), respectively). In fact, in order to make a thorough analysis of the worldwide financial markets we could have taken into consideration every stock market, fixed income index and commodity that is currently traded in the planet. By doing this, we would certainly be able to see how the worldwide resources move along what we understand as global markets. However, the difficulty of doing so would be prohibitively expensive in computational cost, so a simplification has had to be made.

The approach we have taken consists in considering the major financial centres of the world. This


Figure 1: World map with the main areas that have been studied through the selected market indexes.
simplification is equivalent to assume that the evolution of the worldwide financial markets can be modeled through the stock exchanges with higher trading volume on the planet. Even though such assumption may not be strictly true, it seems a good simplification of the reality, since any major change in the evolution of any market in the world will certainly affect these major financial $h u b s$, so it will be perceived somehow in our model. In Figure 1 we present the countries that are directly studied through the selected financial indexes. A list containing the 19 indexes is presented in Table 3.

The information of the indexes has been converted into one single currency, USD, in order to have a trustworthy representation of the price of the indexes. Such transformation has been done through the spot currency exchange (value for the currency exchange in the foreign exchange market at the end of the week). In case we did not do it, the information provided by the indexes would be partial, since the actual value of the index would depend on the exchange currency rate (which varies every day). Once this dollarization has been carried out, we can discuss the compositional nature of the indexes. The key argument supporting the fact that the information of the indexes is compositional is that in case these indexes were multiplied by a factor $k$, the information would remain unchanged. In other words, the actual numeric value of an index is irrelevant by itself, the important information is how it performs with respect to the rest of indexes. However, even though the set of values for the 19 indexes can be understood as a vector of the simplex, the sum is not constant, so our compositional approach will slightly differ from the classical compositional analysis, thoroughly explained in Aitchison (1986) and Aitchison et al. (2002); provided that the actual sum of the components is unknown and variable we will not be able to define the closure operation. Anyway, the fact of not being able to perform the closure operation, which requires to know the actual value for the sum, will not be necessary, since we will be exclusively interested in the evolution of the indexes with respect to the rest.

It is especially relevant to highlight that since the sum of the indexes is variable and unknown we will not be able to predict actual numeric values for the indexes. From a traditional point of view where the most relevant information of the model is the predicted value for every index, this may seem disappointing; it is certainly not, because in case such algorithm was used for trading purposes, with the information on the relative expected behaviour of the indexes we would be able to get a position in the markets.

The compositional approach has consisted in defining a sequential binary partition and computing the corresponding ILR coordinates, as it was developed in Egozcue et al. (2003) and Egozcue and Pawlowsky-Glahn (2005). The definition of such partition is not random; in fact, according to the information that we had a priori, we have created a partition with several ratios that are of

Table 3: List of the 19 indexes that have been considered for the study.

* In fact, the index we have used is not explicitly the 10 year Hong Kong bond, we have used an index containing a group of Asian bonds issued in US dollars. However, the evolution of such index can be understood as the evolution of the fixed income in that region, though not being strictly Chinese.

| Name of the index | Location of the market | Currency of issue |
| :---: | :---: | :---: |
| Dow Jones Industrial Average (DJIA) | New York, USA | US Dollar (USD) |
| Eurostoxx50 (EUSTX) | Frankfurt, Germany | Euro (EUR) |
| Milano Italia Borsa (MIB) | Milano, Italy | Euro (EUR) |
| Footsei100 (FTSE) | London, UK | Great Britain Pound (GBP) |
| Nikkei225 (NIKKEI) | Tokyo, Japan | Japanese Yen (JPY) |
| Hang Seng Index (HSI) | Hong Kong, Hong Kong | Hong Kong Dollar (HKD) |
| Bovespa | Sao Paulo, Brazil | Brazilian Real (BRL) |
| Australian Stock Exchange (ASX) | Sydney, Australia | Australian Dollar (AUD) |
| 10 year bond USA | USA | US dollar (USD) |
| 10 year bond Japan | Japan | Japanese Yen (JPY) |
| 10 year bond Germany (BUND) | Germany | Euro (EUR) |
| 10 year bond Spain | Spain | Euro (EUR) |
| 10 year bond Hong Kong * | Hong Kong | US Dollar (USD) |
| 10 year bond UK | UK | Great Britain Pound (GBP) |
| 10 year bond Australia | Australia | US Dollar (USD) |
| Gold (futures due earliest date) | New York (NYMEX) | US Dollar (USD) |
| Brent (futures due earliest date) | London (ICE) | US Dollar (USD) |
| Cocoa (futures due earliest date) | London (ICE) | US Dollar (USD) |
| Corn (futures due earliest date) | London (ICE) |  |

interest. This partition is presented schematically in Figure 2. Before moving onto the definition of the coordinates, a few remarks have to be done regarding why we have considered such divisions. First, we have distinguished between two types of indexes: in one group we have the traditional, highly liquid indexes, which are present nearly in every portfolio of any investment firm of the world. These indexes include the high quality sovereign bonds (bonds issued by the US, Germany, the UK or Japan) as well as the most liquid stock market indexes (New York, London, Frankfurt, Japan and Hong Kong). We have also included in this group the Brent and Gold futures because they represent highly liquid assets that are very common as well in any investment portfolio. In the other group we have included the rest of indexes, including Southern Europe assets, as well as Emerging Markets indicators, such as the Sao Paulo Stock Exchange. It is especially relevant to point out that even though alternative divisions based on, for example, the location of the indexes, would have been correct as well, they would not have represented this clear binomial between high quality assets versus riskier assets, which is the one assumed to be closely related to the Google searching pattern. The underlying hypothesis under this procedure is that investors are risk averse: in other words, risky assets will decrease in value when uncertainty arises because their expected returns will not compensate any more the associated risk that must be taken when investing in them (for further information on risk aversion Arrow (1965) and Pratt (1964)).

Once this first division has been made, we have proceeded similarly for each group. For instance, we have divided the high quality assets group into two subgroups, safer assets versus stock market indexes. After that, we have made subsequent divisions according to the same criterion. In the same way, we have divided the indexes contained in the riskier assets group. Overall, the key message is that we have tried to separate the indexes in such a way that makes sense in terms of the decision that any investor would have to make. For example, in case the perspectives for the future worsen, any rational investor would decide to invest in highly liquid, conventional assets, and liquidate any position in the riskier, more volatile indexes. Therefore, the partition that we have to create has to represent such decision in order to be meaningful.

To sum up, the treatment of the financial indexes through a ILR coordinate transformation has been performed because the nature of such data expresses relative information. In order to do so, a strategic sequential binary partition has been defined and the standard procedure to compute the ILR coordinates that was proposed in Aitchison et al. (2002) has been applied. The information containing the sequential binary partition is presented in Table 4. These computations have been made through Codapack and R software (Comas-Cufí and Thió-Henestrosa, 2011; R Development


Figure 2: Scheme of how the 19 indexes have been separated in order to create the binary partition needed to compute the ILR coordinates. The divisions have been numbered in red so the corresponding ILR coordinate is easy to identify.

Core Team, 2008).
After having defined the corresponding treatments on both data sets, an exploratory analysis has been performed. We present the compositional biplots that have been built in Figure 3. It is especially relevant to point out how both biplots appear to follow a helicoidal pattern: it seems that both the explanatory and response variables evolve in a way where there is a somewhat cyclical pattern. However, there is something that makes the situations over time different from the past. Indeed, the fact of observing a cyclical pattern in economic data is something we might have expected, yet the helicoidal shape of the biplots remains unexplained.

## 3 Regression model

The model that has been built consists in a multiple linear regression, where both the explicative and predicted variable have had a previous compositional treatment. This model has been built with R software. In total, we have predicted 18 ILR coordinates with a linear combination of 15 variables (which are the first 15 principal components of the vector containing the CLR transformation of the searches):

$$
I L R_{j}=\beta_{0 j}+\sum_{i=1}^{i=15} \beta_{i j} x_{i}, \forall j \in[1,18]
$$

The normality on the residuals has been analyzed on the 18 regressions, through carrying out the Kolmogorov-Smirnov test (Massey, 1951), the Anderson-Darling test (Anderson and Darling, 1952) and the Shapiro-Wilk test (Royston, 1982). Since we are dealing with a big database, the KS

Table 4: Sequential binary partition that has been used to create the ILR coordinates for the indexes.

|  | DJIA | NIK | MIIB | EUSTX | ASX | BOV | HSI | FTSE | 10US | 10JAP | 10SPA | 10GER | 10AUS | 10HK | 10UK | Brent | Gold | Corn | Cocoa |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| ILR1 | 1 | 1 | -1 | 1 | -1 | -1 | 1 | 1 | 1 | 1 | -1 | 1 | -1 | -1 | 1 | 1 | 1 | -1 | -1 |
| ILR2 | -1 | -1 | 0 | -1 | 0 | 0 | -1 | -1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | -1 | 1 | 0 | 0 |
| ILR3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | -1 | 0 | 1 | 0 | 0 | -1 | 0 | 1 | 0 | 0 |
| ILR4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | -1 | 0 | 0 |
| ILR5 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ILR6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| ILR7 | 1 | -1 | 0 | 1 | 0 | 0 | -1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| ILR8 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 | 0 | 0 |
| ILR9 | 1 | 0 | 0 | -1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ILR10 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ILR11 | 0 | 1 | 0 | 0 | 0 | 0 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ILR12 | 0 | 0 | 1 | 0 | -1 | -1 | 0 | 0 | 0 | 0 | 1 | 0 | -1 | -1 | 0 | 0 | 0 | 1 | 1 |
| ILR13 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -1 |
| ILR14 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ILR15 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 1 |
| ILR16 | 0 | 0 | 0 | 0 | 1 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| ILR17 | 0 | 0 | 0 | 0 | -1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| ILR18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | 1 | 0 | 0 | 0 | 0 | 0 |


$z_{d i}=U(M\rangle^{r}$

Figure 3: Compositional biplots of the data containing indexes (left) and the data containing the popularity of the 200 words (right), built through Codapack software.

Table 5: Results of the tests checking the normality of the residuals. The criterion to reject the hypothesis on the normality of the residuals is that either the Anderson-Darling (AD) or the Shapiro-Wilk (SW) test provides a p-value below the significance level of 0.05 and 0.10 , respectively. The Kolmogorov-Smirnov (KS) test has proved to be too strict when dealing with large databases, so its result has not been considered. Hypothesis of normal distribution on the residuals rejected for all regressions except ILR7, ILR8 and ILR11.

| coordinate | p-value KS | p-value AD | p-value SW | normality |
| :---: | :---: | :---: | :---: | :---: |
| ILR1 | $2.20 \mathrm{E}-16$ | 0.000 | 0.000 | NO |
| ILR2 | $2.20 \mathrm{E}-16$ | 0.009 | 0.003 | NO |
| ILR3 | $2.20 \mathrm{E}-16$ | 0.052 | 0.057 | NO |
| ILR4 | $2.20 \mathrm{E}-16$ | 0.069 | 0.058 | NO |
| ILR5 | $2.20 \mathrm{E}-16$ | 0.000 | 0.003 | NO |
| ILR6 | $2.20 \mathrm{E}-16$ | 0.000 | 0.000 | NO |
| ILR7 | $2.20 \mathrm{E}-16$ | 0.528 | 0.203 | YES |
| ILR8 | $2.20 \mathrm{E}-16$ | 0.338 | 0.354 | YES |
| ILR9 | $2.20 \mathrm{E}-16$ | 0.003 | 0.000 | NO |
| ILR10 | $2.20 \mathrm{E}-16$ | 0.002 | 0.000 | NO |
| ILR11 | $2.20 \mathrm{E}-16$ | 0.141 | 0.108 | YES |
| ILR12 | $2.20 \mathrm{E}-16$ | 0.032 | 0.121 | NO |
| ILR13 | $2.20 \mathrm{E}-16$ | 0.076 | 0.023 | NO |
| ILR14 | $2.20 \mathrm{E}-16$ | 0.000 | 0.003 | NO |
| ILR15 | $2.20 \mathrm{E}-16$ | 0.122 | 0.073 | NO |
| ILR16 | $2.20 \mathrm{E}-16$ | 0.000 | 0.000 | NO |
| ILR17 | $2.20 \mathrm{E}-16$ | 0.000 | 0.000 | NO |
| ILR18 | $2.20 \mathrm{E}-16$ | 0.098 | 0.002 | NO |

Table 6: Fitting indicators for the regression on the 18 coordinates.

| coordinate | $R^{2}$ | coordinate | $R^{2}$ |
| :---: | :---: | :---: | :---: |
| ILR1 | 0.801 | ILR10 | 0.889 |
| ILR2 | 0.791 | ILR11 | 0.609 |
| ILR3 | 0.950 | ILR12 | 0.933 |
| ILR4 | 0.949 | ILR13 | 0.838 |
| ILR5 | 0.724 | ILR14 | 0.385 |
| ILR6 | 0.844 | ILR15 | 0.872 |
| ILR7 | 0.507 | ILR16 | 0.773 |
| ILR8 | 0.818 | ILR17 | 0.740 |
| ILR9 | 0.845 | ILR18 | 0.660 |

test is not appropriate because it rejects the normality hypothesis easily (for further information on the suitability of the KS test for big databases see Babu and Rao (2004)). The results of the corresponding p-values have been presented in Table 5. We have rejected the hypothesis of normality on the residuals unless both the AD and SW test have provided p-values above 0.05 and 0.10, respectively (limits imposed according to Anderson and Darling (1952) and Royston (1982)). We have concluded that normality on the residuals can be rejected for all the regressions except ilr7, ilr8 and ilr11.

It is relevant to point out a few relevant details with regard to the results of the regression. First, recall that no collinearity may be expected, since the variables come from a principal component decomposition so they are orthogonal by definition. Secondly, no leverage problems have been observed due to the data, by evaluating all the points of the model with the Cook distance, according to Cook (1977) and Kim and Storer (1996). Finally, potential autocorrelation patterns on the residuals have been evaluated through the Durbin-Watson test Durbin and Watson (1950, 1951), and the conclusion has been that no autocorrelation may be expected from the present model.

The fitness of the model has been evaluated through the $R^{2}$ indicator. The values for the $R^{2}$ are presented in Table 6. Provided that we are fitting 554 points with a 15 variable model the results seem highly positive. It would be arguable whether 15 variables are too many, and whether a simplified model would be more appropriate. Indeed, since we are performing 18 regressions and each one requires fitting 16 coefficients, our model, which was aimed to be simple, demands nothing less than 288 coefficients. However, since the database we count on is considerably large the present solution seems adequate.

A relevant conclusion of the results is that the F test we have performed in the 18 regressions
has proven the linear model we have defined makes sense; provided that the p-value for the F test in all 18 cases is $2.2 \cdot 10^{-16}$ we can infer that the modeling of the evolution of the financial indexes through an isometric log-ratio transformation (ILR) is undoubtedly meaningful. Several ideas can be confirmed from such strong results. First, modeling the indexes through a binary partition approach is appropriate. Even though we had thoroughly justified why the relevant information on the indexes is on the relative performance with respect to the rest rather than on their absolute evolution, we confirm that this intuitive idea is correct. Second, we can also confirm that the compositional approach we have taken for the information on the popularity of the words is appropriate as well. Indeed, even though the compositional treatment looked adequate according to the nature of the data, we can confirm that by proceeding this way the model is meaningful. Last, we have also confirmed that a linear model is a correct approximation. In fact, even though both compositional treatments on the variables seemed reasonable, we did not have any information on whether a linear model would be enough, which has been the case.

## 4 Discussion of the results

Even though the 18 regressions we have proposed have turned out to be meaningful, the fitting coefficients $R^{2}$ differ, with values that oscillate between 0.385 and 0.95 . It is especially relevant to address such differences and whether this is a situation we could have expected. There are 3 regressions with a fitting coefficient above 0.93 , and provided that we are adjusting 554 points this is an extremely powerful result. The balances associated to these regressions are the following comparisons: triple A assets versus double A assets, triple A bonds versus Gold and European risky assets versus Emerging Markets assets.

If we take a closer look we can see that it is extremely intuitive that these particular ratios are fitted better than the rest with the data from the Google searches, since they represent relationships between assets that behave systematically in the same way regardless of the time period we consider. Let us consider the example where there is a sharp increase in uncertainty and turmoil arises in the markets. In that case, the AAA assets will recurrently perform better than AA assets, so the ratio between them will increase (unless AAA assets stop being more secure than AA). However, in case we think of another ratio different from the previous ones, the behaviour is not that predictable; for instance, in case turmoil arises it is not clear whether the Nikkei (Tokyo Stock Market) will perform better than the Hang Seng (Hong Kong Stock Market), so it seems reasonable that the $R^{2}$ drops to 0.609 . Therefore, from the results we obtain from the $R^{2}$ we can infer that there are several combinations of assets that recurrently behave in a similar way when similar situations happen, whereas there are other combinations of assets where the behaviour is not easily predictable.

On the other extreme of the fitting, we have the ratio between the Milano Stock Exchange and the 10 year Spanish bond, where the $R^{2}$ is 0.385 . This seems to prove that the popularity of the terms in Google cannot explain the relationship between these two types of assets, or in case it does the fitness is not accurate. This is a result that might seem shocking in the beginning, but it is not. It is widely assumed that there is an inverse correlation between the stock markets and the sovereign bond yields: once uncertainty arises, funds tend to move to safer assets (bonds) and avoid risky assets (stocks) until confidence is restored. Therefore, it would seem sensible to expect a good fitness in the regression, since bonds should systematically behave better than stocks at rough periods and the other way around in case the optimism is restored. However, that is not what happened in Southern Europe in the time period we have studied. In Figure 4 we have presented the evolution of ILR14 (Milano Stock Exchange versus 10 year Spanish bond) and the evolution of ILR4 (AAA assets versus AA assets). If we look closely at the evolution of Milano stocks versus Spanish bonds we will see that there are localized discrepancies between the fitted values and the real values: in other words, it is not that the model consistently fits badly the data, the fact is that there are certain time periods where the prediction is clearly different from the real situation. If we look carefully at these discrepancies, we can easily understand what has happened. The first


Figure 4: Evolution of the ILR14 and ILR4 coordinates, corresponding to the ratios involving the Milano Stock Exchange versus the 10 year Spanish bond yield and the AAA assets performance versus AA assets in the time period 2004-2014. Each year has been marked with a vertical line (each year is equivalent to 52 weeks).
discrepancy arises in 2010, which corresponds to the Eurozone debt crisis: in that situation, the model gets information of uncertainty, and predicts that bonds should perform better than stocks, but this is not the case, since in this period Southern Europe bonds were assumed to be extremely risky as well, so all the funds were escaping both from the stock market and the sovereign bond market.

The second result we can comment on is that despite having different values for the $R^{2}$ on the 18 regressions, the information from the Google browser users is clearly explicative of the behaviour of the relationships amongst different market indicators the following week, even though the knowledge of such users on these market indicators might be (very) limited. In other words, even though the Google users may not be individually able to infer an accurate prediction on how, for example, AAA assets will perform with respect to Gold the subsequent week, they collectively provide an extremely good approximation of such behaviour. This phenomenon was first introduced in Surowiecki (2004) and referred as Wisdom of Crowds (for further information on this concept Surowiecki (2005) and Koohang et al. (2008)).

## 5 Conclusions

The present project has tried to address the potential correlation between the popularity of 200 words in the Google browser and the evolution of a selection of key financial indexes form all over the world. Rather than trying to predict the actual value for the indexes, a compositional approach has been taken so the regression has been made on the ILR coordinates of the indexes, corresponding to the log-ratios of the indexes in the way we have considered most appropriate, according to the risk averse nature of any rational investor. The results of such regressions have turned out to be very positive, since we have obtained that all 18 regressions are meaningful and present good fitness.

The fact of having obtained a meaningful model leads to consider that even though the individual knowledge of the Google users on the relationships of several dollarized worldwide economic indicators may be limited, collective search patterns are an accurate estimator of further evolution of what we understand as global financial markets. The present project has not been aimed to address this phenomenon, recurrently referred as wisdom of crowds, yet it has proven that the information of the popularity of certain words that are looked up by people who not necessarily have thorough financial background is highly correlated with the performance of several economic indexes the subsequent week.

The second conclusion we can highlight is that the best fitness coefficients of the model have been for those groups of assets that recurrently behave in the same way. For example, AAA assets systematically behave better than AA assets during recessions and worse in times of expansions. This fact is consistent with what we could expect from a model with the characteristics we have defined: we have calibrated a model with the assumption that the variable we are trying to predict will behave the same way under similar circumstances, so its fitness will be insufficient in those cases where the relationship between the assets does not evolve in a similar way under similar situations. However, in case we are able to identify groups of assets whose relationship is consistent over time, the fitness of the model becomes extremely powerful, with correlation coefficients above 0.93 .

It is left to discussion whether the present approach is the best that can be taken. Indeed, there is no evidence that any trading activity based on the present model would have performed better than another one based on the regressions on the raw indexes (with any possible modifications such as a log-scale transformations). However, the identification of key ratios whose evolution is consistently well correlated with online search patterns is absolutely positive.

## 6 Acknowledgements

The present research work has been developed during the time that Robert Ortells has benefited of a grant within the program "Becas de Colaboración" of the Spanish Ministerio de Educación, for the development of the project "Correlació entre l'evolució dels mercats financers globals i la popularitat de diferents paraules al buscador Google" during 2014-2015. This project has been developed in the Applied Mathematics III Department of Universitat Politècnica de Catalunya (UPC). This research has been also partially funded by the Spanish Ministerio de Economía y Competitividad under project 'Metrics' Ref.MTM2012-33236, and by the Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) of the Generalitat de Catalunya under the project "Compositional and Spatial Analysis" (COSDA) (Ref: 2014SGR551;2014-2016).

## REFERENCES

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). Journal of the Royal Statistical Society, Series B (Statistical Methodology) 44(2), 139-177.

Aitchison, J. (1986). The statistical analysis of compositional data. Monographs on statistics and applied Probability. Chapman and Hall Limited (reprinted in 2003 with additional material by the Blackburn Press).

Aitchison, J., C. Barceló Vidal, J. Egozcue, and V. Pawlowsky-Glahn (2002). A concise guide for the algebraic-geometric structure of the simplex, the sample space for compositional data analysis.

Anderson, T. and D. Darling (1952). Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes. Annals of Mathematical Statistics 23, 193-212.

Arrow, K. (1965). Aspects of the Theory of Risk Bearing. The Theory of Risk Aversion. Helsinki: Yrjo Jahnssonin Saatio, Reprinted in: Essays in the Theory of Risk Bearing, Markham Publ. Co., Chicago, 1971.

Babu, G. and C. Rao (2004). Goodness-of-fit tests when parameters are estimated. Technometrics (American Statistical Association) 66, 63-74.

Bordino, I. et al. (2012). Web Search Queries Can Predict Stock Market Volumes. PLoS One 7(7), e40014.

Choi, H. and H. Varian (2012). Predicting the Present with Google Trends. The Economic Record 88, 2-9.

Comas-Cufí, M. and S. Thió-Henestrosa (2011, May 2011). CoDaPack 2.0: a stand-alone, multiplatform compositional software. See Egozcue et al. (2011). ISBN 978-84-87867-76-7.

Cook, R. (1977). Detection of influential observations in linear regression. Technometrics (American Statistical Association) 19, 15-18.

Durbin, J. and G. Watson (1950). Testing for serial correlation in least squares regression, I. Biometrika 37, 409-428.

Durbin, J. and G. Watson (1951). Testing for serial correlation in least squares regression, II. Biometrika 38, 159-179.

Egozcue, J. and V. Pawlowsky-Glahn (2005). Groups of parts and their balances in compositional data analysis. Mathematical Geology 37, 7, 799-832.

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

Egozcue, J., R. Tolosana-Delgado, and M. Ortego (Eds.) (2011, May 2011). Proceedings of the 4 th International Workshop on Compositional Data Analysis, Sant Feliu de Guíxols (Girona). Departament de Matemàtica Aplicada III (Universitat Politècnica de Catalunya): CIMNE. ISBN 978-84-87867-76-7.

Fusion Media Limited (consulted in 2014). www.investing.com, Stock Market Quotes and Financial News.

Ginsberg, J. et al. (2009). Detecting influenza epidemics using search engine query data. Nature 457, 1012-1014.

Google Inc. (Consulted in 2014). Google Trends. 1600 Amphitheatre Parkway. Mountain View. CA 94043.

Kim, C. and B. Storer (1996). Reference values for Cook's distance. Communications in Statistics - Simulation and Computation 25, 691-708.

Koohang, A., K. Harman, and J. Britz (2008). Knowledge Management: Theoretical Foundation (chapter 6: Network analysis and crowds of people as sources of new organisational knowledge). Informing Science Press, Santa Rosa, CA, US.

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

Martín-Fernández, J., K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo (2012). Modelbased replacement of rounded zeros in compositional data: classical and robust approach. Computational Statistics 83 Data Analysis 56, 2688-2704.

Massey, F. (1951). The Kolmogorov-Smirnov test for Goodness of Fit. Journal of the American Statistical Association 46(253), 68-78.

Pratt, J. (1964). Risk Aversion in the Small and in the Large. Econometrica 32(1-2), 122-136.
Preis, T., H. Moat, and H. Stanley (2013). Quantifying Trading Behavior in Financial Markets Using Google Trends. Scientific Reports 3, 1684.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Royston, P. (1982). Algorithm AS 181: The W test for normality. Applied Statistics 31, 176-180.
Surowiecki, J. (2004). The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nation. Little, Brown.

Surowiecki, J. (2005). The Wisdom of Crowds. Anchor Books.
Yahoo (Consulted in 2013). Yahoo Finance. 701 First Avenue, Sunnyvale, CA 94089.

# Physical Activity and Health Outcomes Analysed using Compositional Methods <br> J. Palarea-Albaladejo ${ }^{1}$ and S. F. M. Chastin ${ }^{2}$ <br> ${ }^{1}$ Biomathematics \& Statistics Scotland, Edinburgh, UK; javier.palarea@bioss.ac.uk <br> ${ }^{2}$ Glasgow Caledonian University, Glasgow, UK; sebastien.chastin@gcu.ac.uk 


#### Abstract

The course of a day comprises a sequence of periods of sleep, sedentary behaviour (SED), moderate to vigorous activity (MVPA) and light activity (LIPA) incidental to tasks of daily living. These behaviours affect health in different ways. For example, it has been shown that 5 to 7 hours of sleep and over 30 minutes of MVPA are associated to better health outcomes. To date, the time allocated to each of these behaviours and its relationship to health has been studied in isolation or with partial adjustment for time spent in other behaviours. However, the time allocated to different activities over a day is essentially compositional, it represents a multivariate vector of positive amounts subject to a constant sum, a day is 24 hours long. Time spent in one behaviour necessarily displace time spent in, at least, another one. Hence, the relative and symmetric scale of measurement of physical activity patterns must be considered when conducting statistical analyses aimed at investigating relationships between activities, but also relationships between them and health outcomes. Using data from the U. S. National Health and Nutrition Examination Survey (NHANES), in this work we carry out a first compositional approach to investigate the effects of physical activity patterns on biomarkers of health issues such as obesity, diabetes or cardiovascular disease. The results show that the composition of the day is significantly associated with markers of obesity, cardiovascular and diabetes risks. Moreover, the compositional approach provides unique new insights into this field of research.


Keywords: Physical activity; health risks; compositional data; log-ratio analysis.

## 1 Introduction

The course of a day is made of a sequence of periods of sleep, sedentary behaviors (SB) such as watching television, light activity (LIPA) incidental to tasks of daily living and moderate to vigorous activity (MVPA). Research showed that these behaviors are related to health. For example, there is good evidence that 5 to 7 hours sleep (Bansil et al., 2011), and over 30 minutes of MVPA per day (Haskell et al., 2007) are associated with better health outcomes. More recently, it has emerged that time spent in LIPA might play a positive role in preventing obesity (Kotz and Levine, 2005) and that sedentary time is detrimental to health (Owen et al., 2009). To date, the time allocated to each of these behaviors through the day and its relationship to health has been studied in isolation or with only partial adjustment for time spent in other behaviors (Pedisic, 2014). However, very little attention has been paid to the combined effect of the behavioral time budget as a whole and the technical peculiarities of this type of data. Indeed time is finite, in a circadian or diurnal cycle, and time spent in one behavior necessarily displace time spent in, at least, another one. This implies that the measurements are not independent but rather co-dependent relative amounts of time, which together add up to a finite total time equal to either 24 hours, the waking day or wear-time depending on the measurement protocol. That is, they are constrained data containing only relative information. This, for example, explains why standard statistical analyses commonly face collinearity problems when dealing with this type of data. Besides, the usual correlation measure becomes an inconsistent measure of pair-wise relationships (Pearson, 1897). It is also a sensible requirement that inferences about some behaviors should not depend on the presence or absence of any other behaviors in the data. Moreover, conclusions should not depend either on whether the data are expressed in raw units, say minutes, or re-scaled to e.g. proportions or percentages.

These particularities can be accommodated when the entire profile of physical behaviors is analyzed within a compositional data analysis framework based on log-ratios of the original variables (Aitchison 1986, Pawlowsky-Glahn and Buccianti, 2011). The distribution of time spent in physical behaviors is intrinsically compositional. The aim of this study is to introduce a compositional approach to the most common data analysis in physical activity behavior epidemiology using information from a well-known database and demonstrate its benefits.

## 2 Study design and data analysis

This study considers data from the 2005-6 cycle of the National Health and Nutrition Examination Survey (NHANES) designed to assess the health and nutritional status of a representative sample of the United States population. This database is well-characterized and has been extensively used to determine the association of daily time spent sleeping, sedentary and in light and moderate to vigorous activity with cardio-metabolic health markers. The 2005-6 cycle is the only cycle that included concurrent waking day assessment of activity via accelerometry and data on sleep duration. After applying inclusion criteria based on validity of accelerometry data and non-missing self-reported sleep, covariates and biomarker data, a sample of 1937 adult ( 21 to 64 years old inclusive) participants was available for analysis.

### 2.1 Data

The day was partitioned in proportion of time spent in four physical behaviors: sleep, sedentary behavior (SB), light activity (LIPA) and moderate to vigorous activity (MVPA). Sleep duration was self-reported as an integer from 1 to 24 hours and then expressed as a proportion of 24 hours. Time spent in SB, LIPA and MVPA was assessed objectively using an accelerometer worn for seven days during waking hours. Days when the accelerometer was worn for at least 10 hours were considered valid. Participants were included if at least 5 valid days of their activity were available. Minutes spent in each of these three behaviors were tallied per day and averaged over all available valid days. The final values of SB, LIPA and MVPA were expressed as proportions of 24 hours and combined with sleep time to obtain the composition of the day for each participant in proportions adding up to one.

As to cardio-metabolic markers, Body Mass Index (BMI), waist circumference and blood pressure were measured. Blood samples were analysed for non-fasting high-density lipoprotein (HDL) cholesterol and Creactive protein (CRP). A sub-sample also provided fasting specimen for measurement of fasting low-density-lipoprotein (LDL) cholesterol, triglycerides (TRI), plasma glucose (GLU) and insulin (INS). The latter were used to compute homeostasis model assessment of insulin resistance (HOMA).

Potential confounding factors were extracted from self-reported questionnaires. These included sociodemographic variables; age, sex, race/ethnicity (Mexican American, other Hispanic, non-Hispanic white, non-Hispanic black, or other/multi-racial); marital status (married/living together, widowed, divorced, separated, or never married); education ( $<12$ years, 12 years or equivalent, or some college or above); work status (not working, working part-time ( $<35$ hours/week), or working full-time ( $\geq 35$ hours/week); and ratio of family income to poverty level. Lifestyle factors considered were smoking status, consumption of caffeine and alcohol and total energy and saturated fat dietary intake. Self-reported health (SF-12), previous diagnosis of cancer, diabetes, cardiovascular disease and stroke diabetes; and current use of diabetic, antihypertensive or lipideamic drugs were also extracted.

### 2.2 Statistical analysis

Firstly, well-known compositional descriptive statistics of central position and dispersion were computed for the physical behavior data. Namely, we used the compositional geometric mean and the variation matrix. This latter contains all the pair-wise log-ratio variances between parts of the composition. These measures are coherent with the relative and symmetric scale of the data and have been shown to be optimal representatives of, respectively, the center and the variability of a cloud of compositional data points (Aitchison, 1986; Pawlowsky-Glahn and Egozcue, 2002).

Linear regression models were fitted to examine the associations between cardio-metabolic risk markers and time spent in sleep, SB, LIPA and MVPA. The physical activity composition acted as explanatory variable and a health outcome as the response variable. The analysis was based on isometric log-ratio (ilr) transformations (Egozcue et al., 2003). Denoting a $d$-part physical behavior composition by $B=$ $\left[b_{1}, \ldots, b_{d}\right]$, where the components or parts $b_{i}$ are positive amounts satisfying $\sum_{i=1}^{d} b_{i}=k$, with $k$ being the constant total (e.g. 100 if working on percentages), one ilr transformation particularly convenient for regression analysis computes the ilr-coordinates $z_{i}$ as

$$
\begin{equation*}
z_{i}=\sqrt{\frac{d-i}{d-i+1}} \ln \frac{b_{i}}{\sqrt[d-i]{\prod_{j=i+1}^{d} b_{j}}} \tag{1}
\end{equation*}
$$

with $i=1, \ldots, d-1$ (Hron et al. 2012). The regression equation modelling the health outcome $(Y)$ on the ilr-transformed coordinates and potential covariates adopts the form

$$
\begin{equation*}
Y=\gamma_{0}+\gamma_{1} z_{1}+\gamma_{2} z_{2}+\gamma_{3} z_{3}+\cdots+\gamma_{d-1} z_{d-1}+\text { covariates }+\varepsilon, \tag{2}
\end{equation*}
$$

being $\varepsilon$ a normally distributed random error term. This model accounts for the combined effect of all parts of the composition. It does not matter in what order the parts are transformed; the model produces equivalent fitting in all cases with identical $R^{2}$ coefficient, residuals and estimated intercept $\gamma_{0}$. However, the interpretation of the $\gamma_{i}$ coefficients requires some considerations. In standard linear regression we would normally interpret regression coefficients as the strength of the association between the behavior $b_{i}$ and $Y$. This is often used to understand if time spent in a specific behavior has an independent effect on $Y$ and to quantify this potential effect. However, because of the compositional nature of the time spent in physical behaviors, this way of thinking is largely nonsensical and should be abandoned. Instead we should reason in terms of relative amount spent in one behavior with respect to the others. From Equation (1), it is easy to see that the ilr-coordinate $z_{1}$ basically informs about the ratio between time spent in behavior $b_{1}$ and all the others. Therefore $\gamma_{1}$ can be directly interpreted as the strength of the association between the amount of time spent in $b_{1}$ relative to the other behaviors and the outcome $Y$. The p -value associated to $\gamma_{1}$ can be used as usually to determine whether the behavior $b_{1}$ is statistically significant or not to explain the variation in $Y$. It may give a hint that $b_{1}$ is a significant part of the composition for $Y$. However, it should not be interpreted as a sign that $b_{1}$ is a predictor of $Y$ independently of the other behaviors, or that $b_{1}$ is independently associated with $Y$. It is not possible though to interpret $z_{2}$, or generally $z_{2}, \ldots, z_{d-1}$, in the same way using the same model in Equation (2). So the question is how to extract information also about $b_{2}, \ldots, b_{d}$, how to obtain the same type of information for the other behaviors. Thanks to the permutation invariance of the regression fitting mentioned above, it is possible to construct $d$ equivalent models with each behavior sequentially playing the role of first part of the composition and being transformed into $z_{1}$, and then interpret each $\gamma_{1}$ and its associated $p$-value sequentially:

$$
\left\{\begin{array}{l}
Y=\gamma_{0}+\gamma_{1}{ }^{1} z_{1}{ }^{1}+\cdots+\gamma_{d-1}{ }^{1} z_{d-1}{ }^{1}+\text { covariates }+\varepsilon  \tag{3}\\
\vdots \\
Y=\gamma_{0}+\gamma_{1}{ }^{d} z_{1}{ }^{d}+\cdots+\gamma_{d-1}{ }^{d} z_{d-1}^{d}+\text { covariates }+\varepsilon
\end{array}\right.
$$

A superscript $i$ indicates that the behavior $b_{i}$ is used in the numerator of Equation (1) to built $z_{1}$. We then need to interpret each $\gamma_{1}{ }^{i}$ from Equation (3) which gives information about the association between the outcome $Y$ and the relative amount of time spent on the behavior $b_{i}$. These models were used to estimate: (a) the association of each behavior with the outcome adjusted for the time spent in all and each of the other behaviors, (b) the combined effect of the relative distribution of all the behaviors and (c) the effect of displacing time from one behavior to another one.

We are generally interested in quantifying the size of the effect of the behaviours on health outcomes. In traditional regression analysis, the regression coefficients are directly interpreted as the change in $Y$ associated with a change in $b_{i}$. For compositional data this is not the case because a change in $b_{i}$ is necessarily also concomitant with a change in the other behaviors. Moreover, the coefficients $\gamma_{i}$ associated to the log-ratios coordinates $z_{i}$ are not easily interpretable in terms of units of change of the raw behaviors. However, it is possible to quantify the effect of changing $b_{i}$ by considering log-ratios between time spent in $b_{i}$ and the other behaviors. This has the advantage of, not only giving a quantification of the effect of changing time spent in $b_{i}$, but also quantifying this depending on which other behavior this change in $b_{i}$ is displacing. This can be achieved by computing a change matrix by firstly applying the inverse ilr transformation to the regression coefficients $\gamma_{1}, \ldots, \gamma_{d-1}$ of the linear model [Eq. (2)] to obtain the composition $U=\left(u_{1}, \ldots ., u_{d}\right)$ associated to them. Then, the change matrix between behaviors, of elements $c_{i j}$, is given by

|  | $b_{1}$ | $b_{2}$ | $\ldots$ |  |
| :---: | :---: | :---: | :---: | :---: |
| $b_{1}$ | 0 | $c_{12}=\frac{\ln \left(u_{1} / u_{2}\right)}{4 d^{2}}$ | $\cdots$ | $c_{1 p}=\frac{\ln \left(u_{1} / u_{d}\right)}{4 d^{2}}$ |
| $b_{2}$ | $c_{21}=\frac{\ln \left(u_{2} / u_{1}\right)}{4 d^{2}}$ | 0 | $\ldots$ | $c_{2 p}=\frac{\ln \left(u_{2} / u_{d}\right)}{4 d^{2}}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | 0 | $\vdots$ |
| $b_{d}$ | $c_{p 1}=\frac{\ln \left(u_{d} / u_{1}\right)}{4 d^{2}}$ | $c_{p 2}=\frac{\ln \left(u_{d} / u_{2}\right)}{4 d^{2}}$ | $\ldots$ | 0 |

A value $c_{i j}$ provides information about the change in $Y$ in response to change in the ratio between the times spent on behaviors $b_{i}$ and $b_{j}$ by the mathematical constant $e$, which is approximately equal to 2.718 . If $c_{i j}$ is positive the change corresponds to an increase in $Y$. That is, this matrix informs about the effects on the outcome of all possible substitutions between behaviors in the composition of the day. More directly relevant and interpretable $c_{i j}$ values could be obtained if, for example, we could evaluate the effect of substituting 10 minutes and compute the change in log-ratio for all pairs of behaviors around a set point, say the compositional center. Dividing this change by $c_{i j}$ would give the estimated effect on $Y$ of substituting 10 minutes of $b_{j}$ for $b_{i}$. Note also that the values $c_{i j}$ of the change matrix give a trustworthy estimation of what pairs of behaviors are more influential on the outcome. High values of $c_{i j}$ indeed indicate that the two behaviors $b_{i}$ and $b_{j}$ involved are important as a ratio.

Finally, note that non-normally distributed outcome variables were log-transformed. Confounders were entered in the models as covariates by backward elimination and were retained if the corresponding pvalues were $<0.2$. The same set of confounders was considered for each outcome. The linearity of the association between predictors and outcome, as well as the usual requirements for model residuals, were examined. Statistical significance was concluded for p -values below the usual 0.05 significance level. All analyses were conducted using the R statistical system version 3.1.1.

## 3 Results

Descriptive statistics of the percentage time spent on the four behaviors obtained are displayed in Table 1. On average, most time is spent on sedentary behavior, whereas least relative time is, by far, spent on MVPA. The data variability is summarized in the variation matrix. A value close to zero implies that the times spent
on the two behaviors involved in the ratio (arranged by rows and columns) are highly proportional. For example, the variance of $\ln ($ sleep/SB) is 0.148 , which reflects the highest (proportional) relationship or codependence between any two behaviors. On the other end, it can be observed that the highest log-ratio variances all involve MVPA, which shows that time spent in MVPA is the least co-dependent on the other behaviors. This could explain why the effect of other behaviors can appear spuriously independent of MVPA using standard statistics.

Table 1. Compositional descriptive measures of the percentage time spent on sleep, sedentary behavior (SB), light activity (LIPA) and moderate to vigorous activity (MVPA): compositional center and variation matrix.

|  | Sleep | SB | LIPA | MVPA |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Compositional center |  | 28.71 | 40.58 | 29.23 | 1.48 |
|  | Sleep | 0 | 0.148 | 0.168 | 1.077 |
| Variation | SB | 0.148 | 0 | 0.248 | 1.285 |
| matrix | LIPA | 0.168 | 0.248 | 0 | 0.909 |
|  | MVPA | 1.077 | 1.285 | 0.909 | 0 |

The distribution of the sample compositions is shown in Figure 1 by means of a matrix of ternary plots with three behaviors represented at a time. The overlapped heatmaps represent the distribution of the data points (the more intense the color is the higher the concentration of data points). The dispersion structure of the data is represented by $99 \%$ and $95 \%$ normal-based probability regions around the compositional center. These reflect that the highest variability is found in the direction of the MVPA behavior.


Figure 1: Ternary plots of the compositions of time spent in sleep, sedentary behavior (SB), light activity (LIPA) and moderate to vigorous activity (MVPA) (black dots). The dotted lines refer to $95 \%$ and $99 \%$ normal-based probability regions.

Results from the compositional linear regression models are presented in Table 2. It summarizes the relevant regression coefficients for each behavior and the associated p-values. They correspond to change in the log-ratio of the given behavior to the others. The models were adjusted for age, gender, ethnicity/race, self-reported health, diagnosis of health conditions, educational level, social economic status, smoking
status, alcohol consumption, total dietary calories intake, fat intake, caffeine intake, usage of blood pressure drugs and diabetic drug. The overall significance of the model is also given (Overall column).

Table 2. Coefficients and associated p-values from the compositional regression models of cardio-metabolic markers on the daily physical behavior composition: sleep, sedentary behavior (SB), light activity (LIPA) and moderate to vigorous activity (MVPA). Statistically significant coefficients ( $\mathrm{p}<0.05$ ) are highlighted in bold.

|  | Sleep |  | SB |  | LIPA |  | MVPA |  | Overall |
| :--- | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Marker | $\gamma_{1}{ }^{1}$ | p | $\gamma_{1}{ }^{2}$ | p | $\gamma_{1}{ }^{3}$ | p | $\gamma_{1}{ }^{4}$ | p | p |
| BMI | $\mathbf{- 1 . 4 0}$ | 0.009 | $\mathbf{1 . 4 0}$ | 0.002 | $\mathbf{0 . 9 8}$ | 0.029 | $\mathbf{- 0 . 9 8}$ | $<0.001$ | $<0.001$ |
| Waist cir. | -0.41 | 0.749 | $\mathbf{2 . 4 1}$ | 0.023 | 0.96 | 0.368 | $\mathbf{- 2 . 9 6}$ | $<0.001$ | $<0.001$ |
| Systolic | -1.23 | 0.404 | -0.04 | 0.972 | 1.30 | 0.301 | -0.02 | 0.966 | $<0.001$ |
| Blood Pr. |  |  |  |  |  |  |  |  |  |
| Diastolic | $\mathbf{- 2 . 2 2}$ | 0.048 | 1.03 | 0.269 | 0.70 | 0.461 | 0.49 | 0.198 | 0.003 |
| Blood Pr. | 0.02 | 0.286 | -0.03 | 0.096 | -0.01 | 0.549 | $\mathbf{0 . 0 2}$ | 0.012 | 0.121 |
| HDL $^{\dagger}$ | 0.18 | 0.131 | -0.04 | 0.664 | -0.10 | 0.323 | -0.03 | 0.410 | 0.434 |
| LDL $^{\text {Triglyceri }}$ | $\mathbf{0 . 1 7}$ | 0.018 | 0.09 | 0.132 | $\mathbf{- 0 . 2 1}$ | $<0.001$ | -0.04 | 0.062 | $<0.001$ |
| des $^{\dagger}$ |  |  |  |  |  |  |  |  |  |
| CRP $^{\dagger}$ | $\mathbf{0 . 2 3}$ | 0.024 | -0.04 | 0.562 | -0.06 | 0.472 | $\mathbf{- 0 . 1 2}$ | $<0.001$ | $<0.001$ |
| Plasma $^{\text {Glucose }}{ }^{\dagger}$ | $\mathbf{0 . 0 4}$ | 0.044 | -0.03 | 0.070 | -0.01 | 0.691 | -0.01 | 0.811 | $<0.001$ |
| Plasma $^{\dagger}$ | 0.05 | 0.488 | 0.10 | 0.077 | $\mathbf{- 0 . 1 3}$ | 0.033 | -0.02 | 0.316 | $<0.001$ |
| Insulin $^{\dagger}$ |  |  |  |  |  |  |  |  |  |
| HOMA $^{\dagger}$ | 0.10 | 0.158 | 0.07 | 0.249 | $\mathbf{- 0 . 1 5}$ | 0.020 | -0.03 | 0.295 | $<0.001$ |

${ }^{\dagger}$ Log transformed outcome.
Table 2 shows that the relative distribution of time amongst the four behaviors as a whole is statistically significantly associated with all outcomes except for cholesterol levels (HDL, $\mathrm{p}=0.121$, and LDL, $\mathrm{p}=$ 0.434 ). For each behavior, the regression coefficients enable to estimate their relative effect (compared to the others) when the model is adjusted for the entire distribution of time. For BMI, the proportion of time spent in each behavior compared to the other three was statistically significantly detrimentally associated with SB $\left(\gamma_{1}{ }^{2}=1.40, \mathrm{p}=0.002\right)$ and LIPA $\left(\gamma_{1}{ }^{3}=0.98, \mathrm{p}=0.029\right)$, but favorably with sleep $\left(\gamma_{1}{ }^{1}=\right.$ $-1.40, \mathrm{p}=0.009$ ) and MVPA ( $\gamma_{1}{ }^{4}=-0.98, \mathrm{p}<0.001$ ). Interestingly, the same effect sizes but with opposite effects were observed for SB and sleep. However, for waist circumference only SB $\left(\gamma_{1}{ }^{2}=2.41\right.$, $\mathrm{p}=0.023$ ) and MVPA ( $\gamma_{1}{ }^{4}=-2.96, \mathrm{p}<0.001$ ) showed statistical significance with opposite effect and similar strength. No parts of the composition were statistically significantly associated with systolic blood pressure, but sleep was favorably associated with diastolic blood pressure ( $\gamma_{1}{ }^{1}=-2.22, \mathrm{p}=0.048$ ). The contrast of time spent in MVPA was associated beneficially with HDL levels ( $\gamma_{1}{ }^{4}=0.02, \mathrm{p}=0.012$ ), and this was the only part of the composition exhibiting such result. Statistical evidence of association between MVPA and lower CRP levels was also obtained ( $\gamma_{1}{ }^{4}=-0.12, \mathrm{p}<0.001$ ), although for this outcome sleep resulted statistically significantly associated with higher CRP ( $\gamma_{1}{ }^{1}=0.23, p=0.024$ ). Finally, statistical support was obtained to conclude that the contrast between LIPA and the other behaviors was favorably associated with triglycerides $\left(\gamma_{1}^{3}=-0.21, \mathrm{p}<0.001\right)$, plasma insulin $\left(\gamma_{1}^{3}=-0.13, \mathrm{p}=0.033\right)$, and HOMA ( $\gamma_{1}{ }^{3}=-0.15, \mathrm{p}=0.020$ ); while SB was associated with worse outcomes only for obesity markers; and sleep deleteriously associated with triglycerides $\left(\gamma_{1}{ }^{1}=0.17, \mathrm{p}=0.018\right), \operatorname{CRP}\left(\gamma_{1}{ }^{1}=0.23, \mathrm{p}=0.024\right)$ and plasma glucose ( $\gamma_{1}{ }^{1}=0.04, \mathrm{p}=0.044$ ).

### 3.1 Effect of time re-allocation

In order to further understand the role played by the relative amount of time spent on each behavior making up the composition of the day, we estimated the effect on the outcome (in percentage change about their mean value in the sample: $\mathrm{BMI}=28.9 \mathrm{~kg} / \mathrm{m}^{2}$, waist circumference $=97.7 \mathrm{~cm}$, systolic blood pressure $=$ 120.1 mmHg , diastolic blood pressure $=70.8 \mathrm{mmHg}, \mathrm{LDL}=3.06 \mathrm{mmol} / \mathrm{L}, \mathrm{HDL}=1.41 \mathrm{mmol} / \mathrm{L}$, triglycerides $=1.65 \mathrm{mmol} / \mathrm{L}, \mathrm{CRP}=0.44 \mathrm{mg} / \mathrm{dL}$, glucose $=5.74 \mathrm{mmol} / \mathrm{L}$, insulin $=69.12 \mathrm{pmol} / \mathrm{L}, \mathrm{HOMA}$ $=3.04$ ) of transferring 10 minutes from one behavior to another one around the average composition (Table 1). We chose 10 minutes because this is the smallest unit of change in activity recognized to have beneficial health effect (see e.g. Haskell et al., 2007). The results are presented in Table 3 for each health outcome.

Table 3. Change matrices showing the effect on each outcome of taking away 10 minutes of time from the behavior in columns and allocating them to the behavior in rows. The effect is computed for time re-allocation around the compositional center and expressed as $\%$ change in the outcome about their mean value in the sample (see text for details). The behaviours are sleep, sedentary behaviour (SB), light activity (LIPA) and moderate to vigorous activity (MVPA).


The effect of re-allocating time from one behavior to another one whilst the other two are kept stable was found to be small and not symmetric. For example, for BMI the largest effect was found when 10 minutes of MVPA were displaced by 10 minutes of SB, this changed BMI by $1.21 \%$. However, the opposite, replacing 10 minutes of SB by 10 minutes of MVPA, only changed BMI by $-0.001 \%$. Sleep had a small positive effect on obesity markers. Small positive effects were also observed on systolic and diastolic blood pressure when displacing LIPA and MVPA with sleep, with larger effects observed when sleep replaces MVPA. Replacing MVPA with sleep changed positively $\log (H D L)$ by $0.09 \%$, but only $0.004 \%$ and $0.003 \%$ changes were observed when sleep replaces LIPA and SB respectively. A similar, but more pronounced, pattern was observed for CRP, with sleep reducing the $\log (C R P)$ by $2.1 \%$ when replacing MVPA, but only by $0.004 \%$ and $0.002 \%$ when displacing LIPA and SB in that order. Finally, sleep showed detrimental effects on LDL, triglycerides, glucose and insulin level and HOMA, which were much stronger when sleep displaced MVPA.

SB was found to lead to higher obesity markers if it displaced any of the other behaviors, but only by sizeable amounts if it replaced MVPA. Likewise, SB only had noticeable detrimental effects when displacing MVPA for HDL, triglycerides, glucose, insulin, and HOMA. Much smaller detrimental effects were observed when SB replaced LIPA or sleep time. Detrimental effects were observed when replacing MVPA and sleep with LIPA for both obesity markers, with stronger magnitude when LIPA replaced MVPA. On the contrary, LIPA had favorable effects on LDL, triglycerides, glucose, insulin and HOMA, substantially more pronounced when replacing sleep than when replacing SB. Finally, re-allocating time to MVPA from the other behaviors only had very small effect, most noticeable on HDL.

## 4 Discussion and final remarks

The fact that daily time is finite implies that time spent on one behavior is naturally co-dependent of the time spent on others and the effect of time spent on one behavior should naturally depend on the composition of the rest of the day. This study provides a statistically sound and comprehensive investigation of the associations between cardio-metabolic health markers and the relative distribution of time spent on different physical behaviors that make up the day using a compositional approach. It inherently deals with this issue of relative time adjustment and shows that, for most cardio-metabolic health markers (BMI, waist circumference, blood pressure, triglycerides, C - reactive protein, plasma insulin and HOMA), the composition of time spent in sleep, SB, LIPA and MVPA matters as a whole.

Statistically significant associations between the proportions of time spent in MVPA and SB and the markers were found (Table 2), although Table 3 shows that the strongest effects are observed when SB displaced MVPA. We obtained that the relative times spent in SB and LIPA were both detrimentally associated with obesity markers, HDL, CRP and blood pressure, while MVPA showed a positive relationship. It is important to note that the strength of the negative relationship was higher with SB than
with LIPA. Table 3 shows that the effect of replacing MVPA with SB was stronger than with LIPA. In consequence we can say that, although LIPA might be detrimental, spending the rest of the day in LIPA might be better than sitting (SB). This conveys a relevant message in terms of public health as, even if people cannot or do not want to engage in MVPA, promoting LIPA may be beneficial. For diabetes risk markers (insulin and HOMA) we obtained evidence suggesting that that substituting sitting time for light everyday activities may contribute to preventing and managing diabetes. Moreover, in our analysis sleep was directly beneficially associated with obesity and blood pressure markers, but could have negative effects if it displaced MVPA.

Importantly, our results do not disagree with current evidence about the benefits of MVPA to health, but they show that this relationship depends on the composition of the rest of the day. A striking result is that the effect of substituting MVPA with other behaviors was not the same if the proportion of time in MVPA was increased or decreased. That is, they were not symmetric. For example, replacing 10 minutes of MVPA with 10 minutes of SB corresponded to a $1.2 \%$ higher BMI, while the opposite, replacing 10 minutes of SB with 10 minutes of MVPA, had a 1000 fold lower effect. Hence, the compositional approach allows to obtain unique insights into the research question and contributes to investigate health risks associated to physical activity in a way that other approaches do not allow.

## Acknowledgements

This work has been partly supported by the Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS). The author would also like to thank K. Hron, R. Tolosana and V. Pawlowsky-Glahn for helpful discussions in the early stages of this work.

## References

Aitchison, J. (1986). The statistical analysis of compositional data. London: Chapman \& Hall (re-edited in 2003 by Blackburn Press).

Bansil, P, Kuklina, EV, Merritt, RK, Yoon, PW. (2011). Associations between sleep disorders, sleep duration, quality of sleep, and hypertension: results from the National Health and Nutrition Examination Survey, 2005 to 2008. Journal of Clinical Hypertension 13, 739-743.

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

Haskell, WL, Lee, I-M, Pate, RR, Powell, KE, Blair, SN, Franklin, BA, et al. (2007). Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. Medicine \& Science in Sports \& Exercise 39, 1423-1434.

Hron, K, Filzmoser, P, Thompson, K. (2012). Linear regression with compositional explanatory variables. Journal of Applied Statistics 39, 1115-1128.

Kotz, CM, Levine, JA. (2005). Role of nonexercise activity thermogenesis (NEAT) in obesity. Minnesota Medicine 88, 54-57.

Owen, N, Bauman, A, Brown, W. (2009). Too much sitting: a novel and important predictor of chronic disease risk? British Journal of Sports Medicine 43, 81-83.

Pawlowsky-Glahn, V, Buccianti, A. (2011). Compositional data analysis: theory and applications. Chichester: John Wiley and Sons.

Pawlowsky-Glahn, V, Egozcue, JJ. (2002). BLUE estimators and compositional data. Mathematical Geology 34(3), 259-274.

Pearson, K. (1897). Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London, 489-502.

Pedisic, Z. (2014). Measurement issues and poor adjustments for physical activity and sleep undermine sedentary behaviour research. Kinesiology 46, 135-146.

## Representation of species composition

PAWLOWSKY-GLAHN ${ }^{1}$, V., MONREAL-PAWLOWSKY ${ }^{2,3}$, T., EGOZCUE ${ }^{4}$, J. J.<br>${ }^{1}$ Dept. Informàtica, Matemàtica Aplicada i Estadística, U. de Girona, Spain (e-mail: vera.pawlowsky@udg.edu)<br>${ }^{2}$ AseZOOría salvajes, Barcelona, Spain; (e-mail: taniamonreal@hotmail.com) ${ }^{3}$ IZVG, Keighley, UK;<br>${ }^{4}$ Dept. Matemàtica Aplicada III, U. Politècnica de Catalunya, Barcelona, Spain (e-mail: juan.jose.egozcue@upc.edu)


#### Abstract

An alternative to the de Finetti ternary diagram for representing variability of species composition is presented. The proposed alternative is based on the Aitchison geometry of the sample space. It allows statistical modelling and analysis without the problems derived from spurious correlation, typical of a standard analysis of proportions. The method is illustrated with data representing the species composition of Free and FAD tuna school sets sampled in the Indian and Atlantic Oceans during the 2002-2008 period by purse seiners.


## Introduction

The difficulties for a visual illustration of the variability of the species composition of sampled sets landed by purse seiners in the Indian and Atlantic Oceans moved Fontenau et al. (2010) to propose the use of ternary plots, named after de Finetti (Cannings and Edwards, 1968; Edwards, 2000), to solve this problem. As summarised by Howarth (1996), there has been an extensive use of the ternary diagram in many different science fields. The ternary diagram has a major drawback in that only three part compositions can be visualised. A similar representation with four parts is possible using a tetrahedron. But the most important drawback is that, using standard statistical methods with proportions is known to lead to spurious, nonsensical results (Pearson, 1897; Chayes, 1960). Thus, it appears as desirable to have tools that allow the visualisation and analysis of a larger number of parts. These tools are available in the framework of compositional data analysis based on the logratio approach. Here they are illustrated with real
data corresponding to species composition of tuna landings. In addition to this representation of species composition, it is shown how different groups can be compared using an ANOVA approach.

## 1 The data

The data were kindly provided by Scientists from the IRD (Institut de Recherche pour le Développement, France - http://www.ird.fr/). They correspond to species composition of tuna landings. It is a subset of the so-called "SPECIES" files, i.e. of the detailed species composition of all the sampled sets collected in the Indian and Atlantic oceans. These data have been used and described in Fontenau et al. (2010) to introduce the de Finetti ternary diagrams to show species composition. They correspond to data collected during the period 2002-2008 on the landing or transshipment operations of the EU and associated purse seiners.

The characteristics of the data are summarised in Table 1. The available

|  | fishing mode |  |  |
| :---: | :---: | :---: | :---: |
|  | N | FAD | BL |
| Atlantic Ocean (AO) | 5784 | 3526 | 2258 |
| Indian Ocean (IO) | 4947 | 2857 | 2090 |

Table 1: Sample size of species composition of tuna landings by ocean and fishing mode. $\mathrm{N}=$ total sample size; $\mathrm{FAD}=$ fish aggregating device; $\mathrm{BL}=$ free schools.
data correspond to 5784 operations in the Atlantic Ocean and 4947 in the Indian Ocean, out of which a part has been obtained using a fish aggregating device (FAD), while another corresponds to free schools (BL). For each ocean and each fishing mode three species of tuna have been recorded, YFT $=$ yellowfin, SKJ = skipjack, and BET = bigeye. The data present a major number of zeros, which distribution is summarised in Table 2. Once the two cases with three zeros have been removed, a graphical representation in a ternary diagram is possible, as can be seen in Figure 1. Zero values appear in the border of the ternary diagram. If a single component is zero, the data-point appears on an edge. More precisely, the data point is placed so that the edge is partitioned into two segments, according to the proportion between the two non-zero species. When there are zero-proportions in two parts, the data point is represented in the vertex corresponding to the only recorded species, independently of the value of the single component: the

| ocean | AO |  | IO |  |
| :---: | :---: | :---: | :---: | :---: |
| assoc | FAD | BL | FAD | BL |
| no zeros | 3238 | 515 | 2452 | 98 |
| zeros only in YFT | 15 | 7 | 4 | 2 |
| zeros only in SKJ | 6 | 428 | 10 | 549 |
| zeros only in BET | 237 | 164 | 340 | 147 |
| zeros in YFT and SKJ | 0 | 0 | 0 | 6 |
| zeros in YFT and BET | 22 | 182 | 19 | 310 |
| zeros in SKJ and BET | 8 | 960 | 32 | 976 |
| zeros in YFT, SKJ and BET | 0 | 2 | 0 | 2 |

Table 2: Number of samples with zero and non-zero counts in species composition of tuna landings (see text for details).
proportion of this species is one. However, Figure 1 may be confusing, as data-points with small proportions in one or two species, appear close or very close to those for which there is a zero proportion. This fact redirects us to a deeper problem concerning the representation of proportions. It is the question of the scale. For a couple of proportions like 0.01 and 0.005 , it is clear that the first doubles the second; consequently, their difference or distance should be important. Compare it to the difference between 0.500 and 0.505 ; in this case, the difference can be considered irrelevant. We say that proportions carry only relative information and that the scale is relative. This relative scale is not shown in a ternary diagram, thus claiming for a change of scale for the representation of proportions. The way out is to use logratio transformations of the data in proportions. They account for the relative scale of compositional data and, at the same time, open up a way for applying standard statistical analysis to logratio transformed data (Pawlowsky-Glahn et al., 2015; Mateu-Figueras et al., 2011; Aitchison, 1986).

## 2 Methods

We base our approach on the fact that compositional data, understood as equivalence classes (Barceló-Vidal et al., 2001), can be represented as proportions in a simplex; that the simplex is then a representant of the sample space, and that it admits a Euclidean space structure (Pawlowsky-Glahn and Egozcue, 2001), which leads to the so called Aitchison geometry. The approach is based on the use of general scale invariant logratios, also called log-contrasts.


Figure 1: Representation in the ternary diagram of proportions of landed species. Indian ocean in blue, Atlantic ocean in red. Data-points on the vertices and edges of the triangle correspond to data with two or one zero-proportions.

Denoting the abundances of $D$ species by $\mathbf{x}=\left[x_{1}, x_{2}, \ldots, x_{D}\right]$, log-contrasts appear as expressions like

$$
\ln \frac{x_{1}^{\alpha_{1}} x_{2}^{\alpha_{2}} \cdots x_{k}^{\alpha_{k}}}{x_{k+1}^{\beta_{1}} x_{k+2}^{\beta_{2}} \cdots x_{D}^{\alpha_{D}}}, \quad \sum_{i=1}^{k} \alpha_{i}=\sum_{j=k+1}^{D} \beta_{j},
$$

where $x_{i}$ can represent proportions of landed species in an arbitrary order, and $D$ is the number of considered species, which in the present example is $D=3$. An important property of such log-contrasts is that their value does not change when the data in proportions is multiplied by any positive constant, e.g. 100 for obtaining percentages, or by an arbitrary positive number when they are expressed in abundances. This is useful in the case of the reference data, which are not given in proportions, but in some nonhomogenised units not adding to a prefixed constant. It does not matter, logcontrasts maintain their values independently of the units in which data are presented; in the present case, only ratios between the abundances of some landed species are relevant. There are special cases of logratios which are useful for the representation of compositional data: (a) simple logratios, like $\ln \left(x_{i} / x_{j}\right),(\mathrm{b})$ centered logratios, used in the centered logratio transformation,
which is given by

$$
\begin{equation*}
\operatorname{clr}\left[x_{1}, x_{2}, \ldots, x_{D}\right]=\left[\ln \frac{x_{1}}{g(\mathbf{x})}, \ln \frac{x_{2}}{g(\mathbf{x})}, \ldots, \ln \frac{x_{D}}{g(\mathbf{x})}\right] \tag{1}
\end{equation*}
$$

with

$$
g(\mathbf{x})=\left(x_{1} x_{2} \cdots x_{D}\right)^{1 / D}=\left(\prod_{i=1}^{D} x_{i}\right)^{1 / D}
$$

and (c) balances, which take the form of a logratio of geometric means of groups of components. In the case of the reference data, with $D=3$, a possible choice of balances is

$$
\begin{equation*}
b_{1}=\sqrt{\frac{2}{3}} \ln \frac{x_{3}}{\left(x_{1} x_{2}\right)^{1 / 2}} \quad, \quad b_{2}=\sqrt{\frac{1}{2}} \ln \frac{x_{1}}{x_{2}}, \tag{2}
\end{equation*}
$$

which will be used as Cartesian coordinates of the composition of the three species considered.

The use of logratios to represent compositional data impede the use of zero-proportions, since logarithms of zero take negative infinite values. When zero values in abundances or proportions are understood as values below a detection limit, it is advisable to replace the zero values using some imputation criteria (Martín-Fernández et al., 2011; Palarea-Albaladejo and MartínFernández, 2008; Palarea-Albaladejo et al., 2007; Martín-Fernández et al., 2003). In the present case, with $D=3$ and a large number of zeros, all types of imputation introduce serious distortions in the data set.

An important exploratory tool in compositional data analysis is the clrbiplot (Aitchison and Greenacre, 2002). It is based on the principal component analysis of compositional data (Aitchison, 1983) and is a compositional version of the Gabriel (1971) biplot for real data. Elementary explanations can be found in Pawlowsky-Glahn et al. (2015) and Thió-Henestrosa et al. (2011). Construction of a clr-biplot starts transforming the data-set into their clr's (Eq. 1); a second step is centering the clr-components to place the origin of axes in the center of the data; a singular value decomposition (SVD) of the centered clr-data is then carried out, leading to orthogonal axes which show maximum variability of the data set; finally, the data and the clr-variables are simultaneously represented in a bi-dimensional plot. The bi-dimensional projection of the data set provided by the clr-biplot has the advantage that the data can have a large number of parts or components, thus performing a dimension reduction with the minimum loss of variability. In the present case, where $D=3$, the biplot will represent $100 \%$ of the variability of the data. Biplots can be scaled in several ways, being the so-called covariance
and form biplots the most used. Covariance biplots scale the directions of the clr-variables proportionally to their standard deviations, and the components of data-points appear in a standardized way. Covariance biplots are adequate to visualise relationships between clr-variables. Form biplots normalise the representation of clr-variables, leaving the data-points so that their inter-distances are the Aitchison distances in the simplex (Aitchison, 1983, 1986; Pawlowsky-Glahn and Egozcue, 2001); thus, they are adequate to study the distribution of data-points.

To start with the analysis of the reference data set of landed species of tuna, the zeros have been substituted all by the value 0.00005 in proportions. This value is arbitrary and is chosen to visualise the effect of this raw replacement in a clr-biplot. Figure 2 shows the covariance biplot where the data-points have been colored according to the appearance of zeros (Table 2). The biplot shows a clear separation of the groups corresponding to samples with zeros, thus demonstrating that replacement of zeros introduces artifacts in the data set. From a compositional perspective, this zero replacement re-


Figure 2: Biplot of data after zero replacement; grouped by categories with zero counts (see text for details). Explained variability 100\%.
spects the ratios between the non-zero parts. In the present case, with $D=3$, the dimension of the data set is 2 and, fixing one logratio, the degrees of freedom of a replacement of one zero is one. This effect is shown in Figure 2, in which those cases with only one replaced zero appear dispersed and separated from data-points with no-zeros (blue points), while those cases where
the replacement is carried out in two components, substitution has two degrees of freedom and the replaced data-points appear aligned (gray, pink, and yellow points). This kind of artifacts due to zero replacements appear with any procedure of zero replacement in a composition with three parts ( $D=3$ ); less rough replacement procedures can reduce the separation of the replaced data-points from the data-points without zeros, and even introduce a dispersion similar to that of data without zeros (Martín-Fernández et al., 2011).

For the subsequent analysis, data with zeros have been omitted. The whole data set is classified by ocean (Indian IO, Atlantic AO) and by fishing mode (FAD, BL). Under this two-way classification, the biplot is computed assigning colors to the four different groups of data points. Figure 3 shows the form (upper panel) and covariance (lower panel) biplot. The form biplot (upper panel) is adequate to show the position of the data points. The length of rays is the projection of a unitary vector in the direction of each clr-variable. As in this case $D=3$, the represented variance is $100 \%$, and the rays appear of equal length and unitary, thus telling that they are perfectly projected on the plane of the two principal components. In cases of larger dimension, a short ray points out that the clr-direction is not well projected on the plane. Looking at the data, it can be observed that the classes clearly overlap, although a small shift to the left of data from AO-FAD relative to IOFAD can be observed. Other classes are almost hidden by these two classes. As the first principal component is approximately a logratio of the geometric mean of YFT and SKJ over BET, the shift to the left of AO-FAD means that FAD procedures are more effective in capturing BET, relative to YFT and SKJ, in the Atlantic Ocean than in the Indian Ocean. However, this is only a visual impression and this kind of conclusions should be confirmed statistically.

The lower panel is a covariance biplot. As $D=3$, not many differences can be seen in comparison to the form biplot (upper panel). In a covariance biplot attention is focussed on the clr-variables represented as rays from the origin (in red). The length of the rays is proportional to the standard deviation of each clr-variable corresponding to the logratio of each landed species over the geometric mean of the other two. In a covariance biplot, the main tool of interpretation relays on the segments linking the extreme points of the rays, called links for short. The length of a link is proportional to the standard deviation of the simple logratio between the species in the respective labels. The links in the lower panel of Figure 3 are quite similar, although the link between $\operatorname{clr}(\mathrm{SKJ})$ and $\operatorname{clr}(\mathrm{YFT})$ is a bit shorter than the link between $c \operatorname{lr}$ (SKJ) and $c \operatorname{clr}(\mathrm{BET})$, or the link between $\operatorname{clr}$ (YFT) and $\operatorname{clr}$ (BET). A very short link would imply association between the two involved parts, suggesting


Figure 3: Biplots of data without zeros. Explained variability 100\%. Upper panel: form biplot by ocean and fishing mode. Lower panel: covariance biplot by ocean and fishing mode.

Table 3: Variation array of data without zeros. Upper triangle contains variances of logratios of species by row and column. Lower triangle shows the mean of the logratios. Total variance is 2.61 .

| species | YFT | SKJ | BET |
| :--- | ---: | ---: | ---: |
| YFT |  | 1.66 | 3.56 |
| SKJ | 1.43 |  | 2.62 |
| BET | -0.91 | -2.34 |  |

that the proportions between landed species is almost proportional across the sample (Egozcue et al., 2013; Lovell et al., 2013, 2015).

Features observed in the covariance biplot of Figure 3 are quantified in the variation array (Aitchison, 1986), shown in Table 3. The three variances of simple logratios in the upper triangle are of the same order of magnitude, indicating there is no special association between parts. Their square roots, the standard deviations, are proportional to the links visualised in the covariance-biplot of Figure 3. Mean values of the logratios are shown in the lower triangle of Table 3. For instance, the sample mean of $\ln (\mathrm{BET} / \mathrm{SKJ})$ is -2.34 which indicates that, in overall mean, the landing abundance of BET is small relative to SKJ landings. Variation arrays by ocean, by fishing mode, or by both, show no significantly different features and are not shown here.

In the form biplot (Figure 3, upper panel), the coordinates of data-points on the first principal axis are approximately proportional to the following log-contrast

$$
\begin{equation*}
\ln \frac{(\mathrm{YFT})^{0.58}(\mathrm{SKJ})^{0.21}}{(\mathrm{BET})^{0.79}}, \tag{3}
\end{equation*}
$$

which corresponds to the projections of rays on the first principal axis. However, the interpretation of such coordinates is not easy, specially when there are more than three species in the data set. Therefore, it is common to use more simple log-contrasts to represent and analyse the compositions. This is done using balances, like those in Equation (2).

Within the Aitchison geometry of the simplex, it is advisable to work on orthonormal coordinates, called isometric logratio (ilr) coordinates (Egozcue et al., 2003) to which all the standard statistical methods, devised for real random variables, can be applied (Mateu-Figueras et al., 2011). Those coordinates, shown in Figure 3 (upper panel) or in Equation (3), are ilrcoordinates generated in the principal component analysis. However, in practice, the coordinates used correspond to a sequential binary partition, as described in (Egozcue et al., 2003; Egozcue and Pawlowsky-Glahn, 2005, 2006; Pawlowsky-Glahn and Egozcue, 2011). The generated ilr-coordinates
are then called balances and have a simpler form (Eq. 2). For the reference data, a three part composition, the balance-coordinates used here are

$$
\begin{equation*}
b_{1}=\sqrt{\frac{2}{3}} \ln \left(\frac{\mathrm{BET}}{(\mathrm{YFT} \cdot \mathrm{SKJ})^{1 / 2}}\right), \tag{4}
\end{equation*}
$$

and

$$
\begin{equation*}
b_{2}=\sqrt{\frac{1}{2}} \ln \left(\frac{\mathrm{YFT}}{\mathrm{SKJ}}\right) . \tag{5}
\end{equation*}
$$

They can be visualised in a dendrogram like the one in Figure 4. These


Figure 4: Dendrogram for data without zeros grouped by ocean and fishing mode.
kind of balance dendrograms (Pawlowsky-Glahn and Egozcue, 2011; ThióHenestrosa et al., 2008) are designed to show graphically and simultaneously: (a) the sequential binary partition used for designing the coordinates; (b) the decomposition of the total variance by coordinates; (c) dispersion of balances using box-plots; (d) mean values of balances; and (e) all these features for different populations or groups. In Figure 4, the horizontal junction from the group (SKJ, YFT) to BET corresponds to the balance $b_{1}$ (Eq. 4); the horizontal junction of SKJ and YFT corresponds to the balance between these two species denoted $b_{2}$ (Eq. 5). Note that in presence of more species the dendrogram would have as many junctions as the number of parts minus one. The length of vertical bars over the junction of each balance is proportional to the variance of the balance, thus constituting a decomposition of
the total variance of the sample. The horizontal junction between groups of parts is used to show the dispersion of the balance. All horizontal junctions represent the same range and are scaled accordingly. The zero of the balance (equality of the numerator and denominator of the balance) is the central point, independently of the length of the junction. The boxplots under the junction visualize the dispersion of the sample balances. In the scale of the horizontal junction, the fulcrum of the vertical bars is the mean of the corresponding sample-balance. This is, when the mean balance is placed at the left side, as is the case in Figure 4, it points out that the parts on the left have greater proportions than the parts placed at the right: it works like a lever in equilibrium i.e. a balance in the plain sense. The whole structure is built up using the whole sample, which corresponds to the black vertical lines. When the sample is divided into different populations the variance decomposition of each sub-population is superimposed, respecting the scales of the horizontal junctions. In this way, both variances and mean of each balance can be visually compared.

In Figure 4 the whole sample of landed species (no zeros) is divided by ocean and by association, giving rise to a classification into four groups. As can be seen, variability within each group is pretty similar and some differences might be significant in the means of the balances. For example, the group of tuna landed in the Atlantic Ocean using FAD (FADAO, red) seems to be different in $b_{1}$, while the group of tuna landed in the Indian Ocean from BL (BLIO, green) shows a more differentiated mean in $b_{2}$. Boxplots for $b_{1}$ and $b_{2}$ are reproduced with more detail in Figure 5 (left and right panels respectively).


Figure 5: Boxplots of balance $b_{1}$ (left) and balance $b_{2}$ (right) (enlarged).

Standard statistical methods can be applied to balance-coordinates, including model adjustment or testing of hypothesis, e.g. testing equality of means or performing an analysis of variance. Figures 6 and 7 represent the data in ilr-coordinates (Eqs. 4 and 5), balances in this case, after removing samples with zeros. Data have been split into four sets, crossing Atlantic Ocean (AO) and Indian Ocean (IO) with the extraction techniques, FAD and BL. A bidimensional normal distribution has been adjusted to the data represented by balances. Some isodensity contours of the normal distribution are shown in these figures. Before carrying out a statistical test on the goodness-of-fit, the normal distribution for the balance-coordinates seems a first option for modelling. This corresponds to a normal distribution on the simplex (Aitchison and Shen, 1980; Aitchison, 1982; Mateu-Figueras et al., 2013; Pawlowsky-Glahn et al., 2015). For the whole sample a low, but


Figure 6: Atlantic ocean data without zeros represented in balance-coordinates ( $b_{1}, b_{2}$ ). Normal contours (probabilities, 0.5, 0.75, 0.90, 0.95) fitted to data. Left panel, FAD-fishing mode; right panel, BL-fishing mode.
significant, correlation coefficient of -0.23 is obtained when comparing $b_{1}$ and $b_{2}$. Separating the four groups, the correlation coefficients were -0.25 for AO-BL, 0.33 for IO-BL, -0.26 for AO-FAD, and -0.07 for IO-FAD, thus indicating a weak linear dependence between the balances in the four groups.

A test on equality of means of each balance $b_{1}$ and $b_{2}$ gave $p$-values $<10^{-4}$, indicating that the hypothesis of equality of means should be rejected. The groups responsible of these significative differences for mean balances are those mentioned when looking at the dendrogram in Figure 4. For balance $b_{2}$ the only mean differing significantly from the others is that of IO-BL.


Figure 7: Indian ocean data without zeros represented in balance-coordinates $\left(b_{1}, b_{2}\right)$. Normal contours (probabilities, 0.5, 0.75, 0.90, 0.95) fitted to data. Left panel, FAD-fishing mode; right panel, BL-fishing mode.

For balance $b_{1}$, the mean differing significatively from the others is AO-FAD, specially when compared with IO-BL. This means that the proportions of the three tuna species landed depend on the ocean and on the fishing mode, although the difference in mean is not strong.

## 3 Conclusions

Representation of a composition of species in ternary diagrams, in this case landed species of tuna, was proposed by Fontenau et al. (2010). Here we show that this representation can be complemented using the exploratory and graphical tools of the compositional data analysis. Ternary diagram representations have two main shortcomings: proportions are not represented in an appropriate scale; and they cannot be generalised to more than four species. logratio analysis overcome these shortcomings. The visual tools proposed are: the compositional biplot, the balance dendrogram and scatterplots of balance coordinates. Compositional biplots simultaneously represent data in ilr-coordinates (form biplot) and allows a quite intuitive visualisation (covariance biplot) of relationships between variables. Balance dendrograms allow comparison of means and variances of balance-coordinates, selected by the analyst, from different populations. They graphically show all the elements for an ANOVA analysis to compare means of balances. Finally, representation of compositions by ilr-coordinates allow to construct all de-
sired scatterplots. Statistical modelling of balances or other ilr-coordinates is reduced to the standard statistical multivariate techniques. Moreover, other sets of log-contrasts can be studied separately using standard techniques. Finally, all these representations can be generalised to a large number of parts. For a number of parts greater than 3, the clr-biplot is an optimal projection; the balance dendrogram grows as a tree; and the selection of interpretable balance-coordinates gets more involved, but feasible. In summary, working on coordinates of the simplex is a powerful tool for the representation and analysis of compositional data.

## Acknowledgements

The authors of Fontenau et al. (2010) kindly provided the data used for illustration. This research has been supported by the Spanish Ministry of Education and Science under projects: 'Ingenio Mathematica (i-MATH)' (Ref. No. CSD2006-00032), and 'CODA-RSS' (Ref. MTM2009-13272), from the Spanish Ministry of Economy and Competitiveness under the project 'METRICS' (Ref. MTM2012-33236); and from the Agència de Gestió d'Ajuts Universitaris $i$ de Recerca of the Generalitat de Catalunya under the project Ref. 2009SGR424.

## References

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). Journal of the Royal Statistical Society, Series B (Statistical Methodology) 44 (2), 139-177.

Aitchison, J. (1983). Principal component analysis of compositional data. Biometrika 70(1), 57-65.

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. London (UK): Chapman \& Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Aitchison, J. and M. Greenacre (2002). Biplots for compositional data. Journal of the Royal Statistical Society, Series C (Applied Statistics) 51 (4), 375-392.

Aitchison, J. and S. M. Shen (1980). Logistic-normal distributions. Some properties and uses. Biometrika 67(2), 261-272.

Barceló-Vidal, C., J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2001). Mathematical foundations of compositional data analysis. In G. Ross (Ed.), Proceedings of IAMG'01 - The VII Annual Conference of the International Association for Mathematical Geology, Cancun (Mex), pp. 20 p.

Cannings, C. and A. W. F. Edwards (1968). Natural selection and the de Finetti diagram. Annals of Human Genetics 31, 421-428.

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

Edwards, A. W. F. (2000). Foundations of Mathematical Genetics, $2^{\text {nd }}$ Edition. Cambridge University Press.

Egozcue, J. J., D. Lovell, and V. Pawlowsky-Glahn (2013). Testing compositional association. In P. F. K. Hron and M. T. (Eds.) (Eds.), In: Proceedings of the 5th Workshop on compositional data analysis, CoDaWork 2013, ISBN: 978-3-200-03103-6. pp 28-36.

Egozcue, J. J. and V. Pawlowsky-Glahn (2005). CoDa-dendrogram: a new exploratory tool. In G. Mateu-Figueras and C. E. Barceló-Vidal (Eds.), Proceedings of CoDaWork'05, The 2nd Compositional Data Analysis Workshop, Girona (E). Universitat de Girona, ISBN 84-8458-222-1, http://ima.udg.es/Activitats/CoDaWork05/.

Egozcue, J. J. and V. Pawlowsky-Glahn (2006). Simplicial geometry for compositional data. In Compositional Data Analysis in the Geosciences: From Theory to Practice, Volume 264 of Special Publications, pp. 145-159. Geological Society, London.

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

Fontenau, A., E. Chassot, S. Ortega-Garcia, A. Delgado de Molina, and N. Bez (2010). On the use of the de Finetti ternary diagrams to show the species composition of free and FAD associated tuna schools in the atlantic and indian oceans. Tropical Tunas 65(2), 546-555.

Gabriel, K. R. (1971). The biplot - graphic display of matrices with application to principal component analysis. Biometrika 58(3), 453-467.

Howarth, R. J. (1996, September). Sources for a History of the Ternary Diagram. The British Journal for the History of Science 29(3), 337-356. http://www.jstor.org/stable/4027737.

Lovell, D., V. Pawlowsky-Glahn, and J. J. Egozcue (2013). Have you got things in proportion? a practical strategy for exploring association in highdimensional compositions. In P. F. K. Hron and M. T. (Eds.) (Eds.), In: Proceedings of the 5th Workshop on compositional data analysis, CoDaWork 2013, ISBN: 978-3-200-03103-6. pp 100-110.

Lovell, D., V. Pawlowsky-Glahn, J. J. Egozcue, S. Marguerat, and J. Bähler (2015). Proportionality: a valid alternative to correlation for relative data. PLOS - Computational Biology. dx.doi.org/10.1101/008417.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2003). Dealing With Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. Mathematical Geology 35(3), 253-278. old refs: MBP03, MB+2003.

Martín-Fernández, J. A., J. Palarea, and R. Olea (2011). Dealing with zeros. See Pawlowsky-Glahn and Buccianti (2011), pp. 43-58. 378 p.

Mateu-Figueras, G., V. Pawlowsky-Glahn, and J. Egozcue (2013). The normal distribution in some constrained sample spaces. SORT - Statistics and Operations Research Transactions 37(1), 29-56.

Mateu-Figueras, G., V. Pawlowsky-Glahn, and J. J. Egozcue (2011). The principle of working on coordinates. See Pawlowsky-Glahn and Buccianti (2011), pp. 31-42. 378 p.

Palarea-Albaladejo, J. and J. Martín-Fernández (2008). A modified EM alralgorithm for replacing rounded zeros in compositional data sets. Computers and Geosciences 34(8), 2233-2251.

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

Pawlowsky-Glahn, V. and A. Buccianti (Eds.) (2011). Compositional Data Analysis: Theory and Applications. John Wiley \& Sons. 378 p.

Pawlowsky-Glahn, V. and J. Egozcue (2011). Exploring Compositional Data with the Coda-Dendrogram. Austrian Journal of Statistics 40 (1 \& 2), 103-113.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment (SERRA) 15(5), 384-398.

Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2015). Modeling and analysis of compositional data. Statistics in practice. John Wiley \& Sons, Chichester UK. 272 pp.

Pearson, K. (1897). Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proceedings of the Royal Society of London LX, 489-502.

Thió-Henestrosa, S., J. Daunis-i Estadella, and C. Barceló-Vidal (2011). Exploratory compositional data analysis using CoDaPack 3D. See Pawlowsky-Glahn and Buccianti (2011), pp. 329-340. 378 p.

Thió-Henestrosa, S., J. J. Egozcue, V. Pawlowsky-Glahn, L. Ó. Kovács, and G. Kovács (2008). Balance-dendrogram. A new routine of CoDaPack. Computer and Geosciences 34(12), 1682-1696.

# Insights on the bio-geochemical processes that control sedimentation in Lake Iznik based on compositional data analysis 

P.A. Roeser ${ }^{1}$, S.O. Franz ${ }^{1}$, and T.Litt ${ }^{1}$<br>${ }^{1}$ Rheinische Friedrich-Wilhelms-Universität Bonn; Steinmann Institut für Geologie, Mineralogie und Paläontologie; Nussallee 8, Bonn, 53115, Germany<br>patricia.roeser@uni-bonn.de


#### Abstract

Exploratory compositional data analysis allows insights to the bio-geochemical and physical processes controlling sedimentation in Lake Iznik (NW, Turkey).

Based on the pattern investigation of $c l r$-biplots from geochemical data, generally three main groups of elements are identified: (a) siliciclastic elements, (b) elements related to the carbonate accumulation; (c) elements related to organic productivity. The first group originates from the detrital input to the lake, i.e. catchment derived material that is transported by rivers as particles. The high variance between the siliciclastic elements and the second group, represented by calcium, indicates that a different process controls the majority of calcium accumulation in the sediments. Most of the calcium reaches the lake in dissolved form, originating from carbonate occurrences in the catchment, and further re-precipitates as carbonate from the water column. The third group, characterized by the total organic carbon content, represents the organic productivity within the lacustrine system. Some elements are unrelated to these three groups, such as $\mathrm{Mn}, \mathrm{Na}, \mathrm{S}, \mathrm{P}$, and therefore other processes than the fluvial transport, carbonate accumulation, or organic production of the lake control the behaviour of such elements. Based on the pattern investigation of clr-biplots of mineral data, we conclude that calcite has a distinct origin from aragonite, as from the siliciclastics. Additionally, kaolinite displays relative higher variance to the other aluminosilicates. One source of kaolinite is the weathering of albite. The evaluation of balance-dendograms reveals that the calcium accumulation, i.e. carbonate precipitation in the lake, retains most of the variability within the data-sets.

A pattern comparison between ternary subcompositions of elements and minerals leads to the conclusion of increased chemical weathering since the middle Holocene, implying increased moist conditions. Such conclusions are of importance for further investigations in which the paleo-environment for the Marmara region is reconstructed for the past 31 ka calBP.


Keywords: aragonite, calcite, weathering, kaolinite, albite

## 1 Introduction

Minerals and geochemical elements, as essential building blocks of sediments and its components, are indispensable for understanding the evolution of a lake and its catchment, and also provide understanding for sedimentation processes. The closed nature of the mineralogical and geochemical sedimentary compositions requires that such data is explored and interpreted using the proper methods.

Lake Iznik (NW, Turkey) is a potential site to reconstruct the paleo-environment in the Marmara region (Franz and others, 2006). In the present study, we apply exploratory compositional data analysis to investigate the sediment profile IZN09/LC2\&LC3 (Fig. 1), which originates from the centre of the lake. The final objective of this study is to gain information about the bio-geochemical processes that control sedimentation within Lake Iznik over time, and the relationship between them. The sediment record dates back to circa 31 ka calBP (Roeser, 2014), and five sediment units (SU V to SU I) have been defined based on changes in lithology and magnetic susceptibility (Roeser and others, 2012). These sediment units represent the time intervals from the Late Pleistocene (V and IV), the transition to the Holocene (III), the early Holocene (II), until circa middle Holocene (I). The Last Glacial Maximum (LGM) - at circa 22ka BP, and the Younger Dryas (YD) event - that marks the beginning of the Holocene, are retained in the profile (Fig. 1).


Figure 1: Overview on the composite sediment profile IZN09/LC2\&LC3. Depth scale on the left is given in meters composite depth. Lithology bar highlights the distribution of monosulfides (black spots) and distinct white (w) or black (b) layers. Exemplary, the ratio $\mathrm{Ca} / \mathrm{Ti}$, an expression of the carbonate accumulation in the sediments, is presented as obtained from Itrax core scans [count ratio] in 1 cm resolution, and as obtained from traditional X-Ray Fluorescence in 5 cm resolution. The right bar highlights previously defined sedimentary units. LGM: last glacial maximum. YD: Younger Dryas event.

## 2 Exploratory Compositional Analysis - a short review

The exploratory statistical analysis of compositional data consists of the following steps: (1) compute the descriptive compositional statistics, (2) center the data for a better visualization in ternary diagrams, (3) investigate the clr-biplot to discover patterns in the data set, (4) define an appropriate representation on orthonormal coordinates, and (5) compute the summary statistics of the coordinates and represent results in a balance dendogram (Pawlowsky-Glahn and others, 2011). Each of these steps is summarized in the following paragraphs.

For compositional data, classical statistical summaries carry no information, due to the spurious correlation effect. Therefore, the compositional summary is used, which consists of the determination of the geometrical mean of the composition, and the investigation of the variation matrix, which describes the dispersion in the compositional data set. It is a tool to analyze the relations between the pairs of variables (geochemical elements or minerals). The total variance gives a measure of global dispersion of the composition. Centering the compositional data, means that a perturbation is applied to shift the data set around its barycenter. For this, the samples are perturbed by the opposite of the geometric mean, g-1. Centering the data set is used for better visualization within the ternary diagrams.

A useful exploratory tool which displays the samples and the related variables at the same time is the clrbiplot, which is a projection of the variability of the data (Tolosana-Delgado, 2012). Like in a principal component analysis, the clr-biplot consists of a singular value decomposition, which in this case is applied to the matrix of centered logratio transformed data (Aitchison, 2003a). The clr-biplot has unique mathematical properties, which are used for interpretation of results. The most striking property is that information is contained mainly in the links, i.e. distances between the apexes, or vertices (Aitchison and Greenacre, 2002). The length of links is proportional to the variance in-between the variables. Furthermore, the angles in-between links in the covariance biplot can give an estimation of the relationship between the variables (Aitchison and Greenacre, 2002). For instance, orthogonal links indicate that the associated logratios might be independent. Additionally, coincident vertices (very small variance in-between links) carry the same, or redundant, information. Whilst for collinear vertices the associated subcomposition is a one dimensional biplot, i.e. a compositional line (Aitchison, 2003b). The biplot does not allow a conclusion to be drawn on a single component, because as the projection refers to a clr logratio transformation (which contains the ratio to the geometric mean of the complete composition) it implies a weight of all the parts and not only a single component.

As a result of the exploratory analysis, or given a priori knowledge of the data set, the investigation of a certain sequential binary partition might become interesting. A graphical representation for the coordinates based on such a partition is the balance dendogram (Egozcue and Pawlowsky-Glahn, 2006; Pawlowsky-Glahn and others, 2011). CoDa balance-dendograms are not dendograms in the 'traditional' sense. They go beyond descriptive statistics, and represent a first step to conceptualizing the data, as the balances are defined according to reasonable interpretation approaches. The graphical display of balance dendograms are represented by structure from which we can read mainly two characteristics for each coordinate: (1) the sample variance, which is given by the 'height' of the vertical line (2) the sample mean, which is indicated by the contact point (Pawlowsky-Glahn and others, 2011).

## 3 Lake Iznik: sediment geochemistry and mineralogy

For the sedimentary profile of Lake Iznik (IZN09/LC2\&LC3) three data sets are used for exploratory compositional analysis: (1) geochemical data obtained with X-Ray fluorescence (XRF) directly on wet core halves (Itrax core scanner); (2) inorganic geochemical data for discrete samples, from traditional XRF together with organic content obtained from an elemental analyser; and (3) mineralogical data obtained from X-Ray diffraction. Whilst the Itrax core scanner allows for the analysis of geochemical information in high resolution, conventional XRF on melt tablets and press pellets offers a wider range of measurable elements. Details on analytical measurements are given elsewhere (Roeser, 2014).

All three data-sets are of a compositional nature, therefore they can be expressed as vectors of positive components which sum up to a total, and each part represents relative information. Geochemistry from Itrax core scanners was obtained in 1 cm resolution, resulting in a sample size of circa 1320. The geochemical data measured with traditional X-Ray fluorescence as well as the mineralogy data were obtained in approximately 5 cm resolution, resulting in a sample size of circa 330 .

In the following sections the compositional statistical summary, clr-biplots and selected ternary compositions are presented for the investigated data-sets. Also, possible binary partitions and associated interpretation approaches for the balances are proposed for selected subcompositions. All calculations were undertaken using Codapack (Comas-Cufí and others, 2011).

### 3.1 Major element geochemistry - Itrax core scanner

A subset of elements was chosen from those measured with the Itrax core data, resulting in the subcomposition ( $\mathrm{Ti}, \mathrm{Al}, \mathrm{Fe}, \mathrm{Si}, \mathrm{Ca}, \mathrm{K}, \mathrm{S}$, and Sr ). These elements were selected according to the following criteria: (a) count rates of minimum 150 counts per 10 seconds; (b) low data scattering ${ }^{1}$, (c) the data-sets contained only a few zeros (less than $0.5 \%$ for each variable). No zero replacement was undertaken for the Itrax data, and therefore the respective samples were ignored.

The $c l r$-biplot of the selected subcompotision (Fig. 2) shows that there is one set of elements that have very small variance to each other, characterizing the siliciclastic sedimentation: $\mathrm{Ti}, \mathrm{Al}, \mathrm{Fe}, \mathrm{Si}$, and K . Whereas the alkaline earth metals Ca and Sr characterize the carbonate sedimentation. Strontium is typically incorporated into the lattice of aragonite. Sulfur belongs to neither of the two observed groups and presents one-dimensional variability in relation to both of them. Moreover, the $c l r$-biplot in Figure 2 highlights that the pre-defined sedimentary units (SU I to SU V) present an intrinsic geochemical differentiation.

Table 1. Variation matrix for the subcomposition ( $\mathrm{Al}, \mathrm{Si}, \mathrm{S}, \mathrm{K}, \mathrm{Ca}, \mathrm{Ti}, \mathrm{Fe}, \mathrm{Sr}$ ) obtained from the Itrax core scanner. Total variance is 1.3559. Values presented in bold typeface highlight smallest variances and values in italic typeface mark intermediate variances.

| $\mathbf{X i} \mid \mathbf{X} \mathbf{j}$ | $\mathbf{S i}$ | $\mathbf{S}$ | $\mathbf{K}$ | $\mathbf{C a}$ | $\mathbf{T i}$ | $\mathbf{F e}$ | $\mathbf{S r}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\mathbf{A l}$ | 0.0454 | 0.842 | 0.0434 | 0.9769 | 0.0419 | 0.0511 | 0.5538 |
| $\mathbf{S i}$ |  | 0.6416 | $\mathbf{0 . 0 0 7 4}$ | 0.7368 | 0.0169 | 0.0244 | 0.3605 |
| $\mathbf{S}$ |  |  | 0.6413 | 0.3721 | 0.6861 | 0.6654 | 0.2663 |
| $\mathbf{K}$ |  |  |  | 0.7627 | $\mathbf{0 . 0 0 7 5}$ | 0.0165 | 0.3798 |
| $\mathbf{C a}$ |  |  |  |  | 0.8385 | 0.8448 | 0.132 |
| $\mathbf{T i}$ |  |  |  |  |  | 0.0123 | 0.4331 |
| $\mathbf{F e}$ |  |  |  |  |  |  | 0.4469 |

A possible sequential binary partition (SBP) for the Itrax data is proposed in Table 2. The balances resulting from the SBP are interpreted as: (b1) importance of carbonate phase relative to detrital fraction; (b2) importance of strontium co-precipitation in carbonate; (b3) distribution of monosulfides against detrital elements, (b4) importance of iron for sulfide precipitation; (b5) distribution of aluminosilicates and clay minerals ( $\mathrm{K}, \mathrm{Al}, \mathrm{Si}$ ) against Ti ; i.e. evaluating if Ti is contained in clay minerals; (b6) distribution of micas and kaolinite against Si (which is possibly influenced by the production of biogenic silica by organisms); and (b7) distribution of micas against kaolinite.

Overall, balances b3 and b4 evaluate the geochemical signature of the early diagenetical overprint on the sedimentation. Note that the interpretation of balances b5 and b6 is not without problem, since Al and Si are elements common to a variety of feldspars and clay minerals. The balance dendogram obtained for the SBP from Table 2 is given in Figure 3.

Table 2. Possible sequential binary partition for the data of subcomposition ( $\mathrm{Al}, \mathrm{Si}, \mathrm{S}, \mathrm{K}, \mathrm{Ca}, \mathrm{Ti}, \mathrm{Fe}, \mathrm{Sr}$ ).

| balance | Al | $\mathbf{S i}$ | $\mathbf{S}$ | $\mathbf{K}$ | $\mathbf{C a}$ | $\mathbf{T i}$ | $\mathbf{F e}$ | $\mathbf{S r}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\mathbf{b 1}$ | -1 | -1 | -1 | -1 | +1 | -1 | -1 | +1 |
| $\mathbf{b 2}$ | 0 | 0 | 0 | 0 | -1 | 0 | 0 | +1 |
| $\mathbf{b 3}$ | -1 | -1 | 1 | -1 | 0 | -1 | 1 | 0 |
| $\mathbf{b 4}$ | 0 | 0 | 1 | 0 | 0 | 0 | -1 | 0 |
| $\mathbf{b 5}$ | -1 | -1 | 0 | -1 | 0 | 1 | 0 | 0 |
| $\mathbf{b 6}$ | -1 | 1 | 0 | -1 | 0 | 0 | 0 | 0 |
| $\mathbf{b 7}$ | 1 | 0 | 0 | -1 | 0 | 0 | 0 | 0 |

[^4]

Figure 2: (1st vs. 2nd) projection of clr-biplot for subcomposition (Al, $\mathrm{Si}, \mathrm{S}, \mathrm{K}, \mathrm{Ca}, \mathrm{Ti}, \mathrm{Fe}, \mathrm{Sr}$ ). Data obtained from the Itrax core scanner. The cumulative explained proportions by the principal components are $\operatorname{PC1}(0.8012), \mathrm{PC} 2(0.9482)$, and $\mathrm{PC} 3(0.9722)$


Figure 3: (A) balance-dendogram of data of subcomposition (Al, $\mathrm{Si}, \mathrm{S}, \mathrm{K}, \mathrm{Ca}, \mathrm{Ti}, \mathrm{Fe}, \mathrm{Sr})$; (B) and (C) zoomed parts of the balance dendogram, related to balances b2 $(\mathrm{Ca}, \mathrm{Sr})$ and b5 (Ti in relation to the fine fraction), respectively

From the obtained balance dendogram the highest variance is given for balance 1, highlighting the distinct nature of carbonate deposition in relation to siliciclastic input (Fig. 3). Strontium accumulation in the sediments has very small variances to that of Ca, except for sediment unit IV (Fig. 3C). Among the siliciclastic elements, Fe retains a high amount of variability. Iron shows somewhat elevated variance to S , and is therefore not primarily controlled by monosulfide precipitation. The siliciclastis $\mathrm{Ti}, \mathrm{Si}, \mathrm{Al}$, and K , present nearly constant ratios for all sedimentary units.

### 3.2 Bulk Geochemistry

From the discrete samples the determined major elements are $\mathrm{C}, \mathrm{S}, \mathrm{N}, \mathrm{Si}, \mathrm{Al}, \mathrm{Fe}, \mathrm{Mn}, \mathrm{Mg}, \mathrm{Ca}, \mathrm{Na}, \mathrm{K}, \mathrm{Ti}$, and P. For the carbon content, total inorganic carbon (TIC) and total organic carbon (TOC) fractions were determined separately with an elemental analyser. From the measured trace elements, the selected elements for CoDa analysis were $\mathrm{V}, \mathrm{Cr}, \mathrm{Co}, \mathrm{Ni}, \mathrm{Cu}, \mathrm{Zn}, \mathrm{Ga}, \mathrm{As}, \mathrm{Rb}, \mathrm{Sr}, \mathrm{Y}, \mathrm{Zr}, \mathrm{Nb}, \mathrm{Ba}, \mathrm{W}, \mathrm{Pb}, \mathrm{Th}$, and U, as for these elements less than $10 \%$ of data is below the detection limit ${ }^{2}$.

Generally, the logratio approach is not suitable if the data set contains spots of missing data or zeros. If the zeros are of the type rounded zeros - which is the case when measurement of a determined element is below the detection limit, imputation techniques can be used to replace those zeros. It is practical to replace them by small, non zero, values. If the data set has less than $10 \%$ (of the complete matrix) of values below zero, a non-parametric replacement is appropriate, and the threshold 0.65 of the detection limit is a reasonable option (Martín-Fernández and others, 2003). This was the case for uranium.

The variation matrix of the bulk geochemistry is not shown herein due to formatting issues. The total variance of geochemical data-set is of 2.2021. Pairs of variables with the smallest variances are specially related to the siliciclastic sedimentation, for example $\mathrm{Al}-\mathrm{K}, \mathrm{Al}-\mathrm{Ti}, \mathrm{Rb}-\mathrm{K}$, or $\mathrm{Zn}-\mathrm{Ti}$.

Generally, three groups of elements are recognized in the clr-biplot in Figure 4 (gray circles): (a) siliciclastic elements, (b) elements related to the carbonate accumulation, and (c) and elements related to the organic productivity. The first group originates from the detrital input to the lake, i.e. catchment derived material that is transported by rivers as particles. The detrital elements have nearly coincident vertices highlighting their immobile geochemical character. These elements are $\mathrm{Si}, \mathrm{Al}, \mathrm{Fe}, \mathrm{K}, \mathrm{Ti}, \mathrm{V}, \mathrm{Cr}$, $\mathrm{Co}, \mathrm{Ni}, \mathrm{Cu}, \mathrm{Zn}, \mathrm{Ga}, \mathrm{Rb}, \mathrm{Y}, \mathrm{Zr}, \mathrm{Nb}, \mathrm{W}$, and Th . The high variance between the siliciclastic elements and the second group, represented by calcium, indicates that a different process controls the majority of calcium accumulation in the sediments. Most of the calcium reaches the lake in dissolved form, originating from carbonate occurrences in the catchment, and further re-precipitates as carbonate from the water column. Elements with low to intermediate variance to Ca are the total inorganic carbon (TIC), and Sr. Both are part of the carbonate mineral lattice, TIC in the form of $\mathrm{CO}_{3}{ }^{++}$, and Sr co-precipitates with aragonite from the water column. The organic group is given by the low to intermediate variances between total organic carbon (TOC), and N. Notably, Mg shows very small variance to the organic group, especially to nitrogen. But Mg also has low variance to $\mathrm{Si}, \mathrm{Al}$ and Fe . Also, U shows small variance to the carbonate and organic groups.

Elements associated to neither of the three identified groups are $\mathrm{S}, \mathrm{Mn}, \mathrm{Na}, \mathrm{P}, \mathrm{As}, \mathrm{Pb}$, and U . Hence, for these elements, the rate of supply, and/or fixation into the sediments is further controlled by additional processes.

For example, generally early diagenetical processes, i.e. shortly after burial, might remobilize determined elements under the new geochemical millieu, forming authigenic minerals. Such new mineral formations can remobilize $\mathrm{S}, \mathrm{Pb}$ and As to sulfides (e.g. arsenopyrite), P to phosphates (e.g. vivianite) and Mn to carbonate (e.g. rhodochrosite). Iron can also be subject to such remobilization processes.

From the behavior in the ternary diagrams (Fig.5), Na displays a distinct pattern between the Holocene (sediment units I and II) and the Pleistocene (sediment units III, IV, and V) (Fig.5A). Strontium and Mn behave distinctly for sedimentary unit IV, which encompasses the Last Glacial Maximum (LGM) (Fig.5B and 5C). Notably, the subcomposition ( $\mathrm{U}, \mathrm{Ti}, \mathrm{Mn}$ ) resembles a linear behavior (Fig.5D). Whilst Mn is geochemically mobile under anoxic conditions, U is mobile under oxic conditions and favorable bound to the carbonate lattices.

[^5]

Figure 4: (1st vs. 2nd) and (1st vs. 3rd) projections of the clr-biplot for composition (TIC, TOC, S, N; $\mathrm{Si}, \mathrm{Al}, \mathrm{Fe}, \mathrm{Mn}, \mathrm{Mg}, \mathrm{Ca}, \mathrm{Na}$, $\mathrm{K}, \mathrm{Ti}, \mathrm{P}, \mathrm{V}, \mathrm{Cr}, \mathrm{Co}, \mathrm{Ni}, \mathrm{Cu}, \mathrm{Zn}, \mathrm{Ga}, \mathrm{As}, \mathrm{Rb}, \mathrm{Sr}, \mathrm{Y}, \mathrm{Zr}, \mathrm{Nb}, \mathrm{Ba}, \mathrm{W}, \mathrm{Pb}, \mathrm{Th}, \mathrm{U}$ ). The cumulative explained proportions are PC 1 (0.6828), PC2(0.7915), PC3(0.8499). Gray circle in the $3^{\text {rd }}$ and $4^{\text {th }}$ quadrant refers to ( $\mathrm{Si}, \mathrm{Al}, \mathrm{Fe}, \mathrm{K}, \mathrm{Ti}, \mathrm{V}, \mathrm{Cr}, \mathrm{Co}, \mathrm{Ni}, \mathrm{Cu}, \mathrm{Zn}, \mathrm{Ga}$, $\mathrm{Rb}, \mathrm{Y}, \mathrm{Zr}, \mathrm{Nb}, \mathrm{W}$, and Th ) - element symbols were omitted from figure for simplicity.


Figure 5: Selected ternary diagrams for the bulk geochemical sediment composition

A meaningful subcomposition was selected to contain a representative element for each of the groups: the detrital deposition $(\mathrm{Ti})$, the carbonate accumulation $(\mathrm{Ca})$, the organic production and preservation $\left(\mathrm{C}_{\text {org }}\right)$, in addition to the elements with moderate to high variance to the three groups. A SBP for the subcomposition ( $\mathrm{C}_{\mathrm{org}}, \mathrm{Na}, \mathrm{Mg}, \mathrm{S}, \mathrm{Ca}, \mathrm{Ti}, \mathrm{Mn}$ ) is given in Table 3. The resulting balances are interpreted as: (b1) Importance of detrital phase against carbonate and organic; (b2) organic control on carbonate deposition; (b3) distribution of Mg co-precipitation with carbonates (high Mg calcite); (b4) importance of organic decomposition by sulfur pathway; (b5) importance of albite in relation to detrital; and (b6) importance of redox processes $(\mathrm{Mn})$ against immobile Ti.

Table 3: Possible binary partition for major element subcomposition ( $\mathrm{C}_{\mathrm{org}}, \mathrm{Na}, \mathrm{Mg}, \mathrm{S}, \mathrm{Ca}, \mathrm{Ti}, \mathrm{Mn}$ )

| balance | $\mathbf{C o r g}_{\text {on }}$ | $\mathbf{N a}$ | $\mathbf{M g}$ | $\mathbf{S}$ | $\mathbf{C a}$ | $\mathbf{T i}$ | $\mathbf{M n}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\mathbf{b 1}$ | +1 | -1 | +1 | +1 | +1 | -1 | -1 |
| $\mathbf{b 2}$ | -1 | 0 | +1 | -1 | +1 | 0 | 0 |
| $\mathbf{b 3}$ | 0 | 0 | -1 | 0 | +1 | 0 | 0 |
| $\mathbf{b 4}$ | +1 | 0 | 0 | -1 | 0 | 0 | 0 |
| $\mathbf{b 5}$ | 0 | +1 | 0 | 0 | 0 | -1 | -1 |
| $\mathbf{b 6}$ | 0 | 0 | 0 | 0 | 0 | +1 | -1 |

Similarly to the balance of Figure 3, the highest variance is retained by balance 1 that evaluates the siliciclastic and the carbonate accumulation, and Ca retains high variability within the data-set. Calcium and Mg also retain high variability when related to the preservation of organic matter. For sediment unit IV there is 'abnormal' high variance related to the carbonate accumulation, and the Mn accumulation, when compared to the further sediment units.


Figure 6: Balance dendogram for subcomposition (TOC, $\mathrm{Na}, \mathrm{Mg}, \mathrm{S}, \mathrm{Ca}, \mathrm{Ti}, \mathrm{Mn}$ )

### 3.3 Mineralogy

The major minerals that compose the bulk mineral assemblage of Lake Iznik sediments are quartz, feldspar (plagioclase), kaolinite/chlorite, muscovite/illite, and the carbonate polymorphs calcite and aragonite (Franz and others, 2006). In the present study, the identified feldspar is the sodium end member of the plagioclase series, albite. The concomitant presence of kaolinite and chlorite (chlinoclore) in the bulk sediment is determined by the typical double-peak appearance allowing for good separation in between them. Illite is per definition a smaller grain size of muscovite (Moore and Reynolds, 1997) and therefore, muscovite/illite is considered one mineral phase.

The seven identified mineral phases were quantified along the profile means rietveld analysis, which results in a closed data vector for each depth. The clr-biplot of the mineralogical data set shows that variability of the information is retained by the carbonates; at the same time as aragonite and calcite also have high variance in between each other. The nearly perpendicular links of $\log$ (aragonite/quartz) and $\log$ (calcite/albite), enlighten that the processes related to the mineral ratios originating the links have low correlation, or are independent. Similar independent processes can be found in the biplot for other ratios, although always in relation to the carbonate minerals.

Table 4: Variation matrix for the subcomposition (quartz, calcite, aragonite, muscovite/illite, chlinoclore, kaolinite, albite). Total variance is of 1.7407 . Values presented in bold typeface highlight smallest variances.

| XilXj | Cal | Ara | M / I | Chli | Kao | Alb |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Quartz | 0.4521 | 1.6841 | $\mathbf{0 . 0 3 3 2}$ | $\mathbf{0 . 0 2 5 8}$ | 0.0887 | $\mathbf{0 . 0 2 8 2}$ |
| Calcite |  | 1.3379 | 0.4627 | 0.4126 | 0.4749 | 0.4702 |
| Aragonite |  |  | 1.6383 | 1.6089 | 1.3863 | 1.6447 |
| Musc. /Illite |  |  |  | $\mathbf{0 . 0 3 2 7}$ | $\mathbf{0 . 0 5 7 9}$ | 0.0865 |
| Chlinoclore |  |  |  |  | $\mathbf{0 . 0 5 9 3}$ | $\mathbf{0 . 0 6 1 1}$ |
| Kaolinite |  |  |  |  |  | 0.1391 |



Figure 7: Compositional clr-biplot for major mineral assemblage. Cumulative explained proportions are PC1 (0.7588), PC2(0.9377), PC3(0.9784).

The overall pattern of compositional lines observed within the ternary diagrams of mineral subcompositions highlights the intrinsic geochemical discrimination of the sedimentary units, especially with units IV and II as extremes (Fig. 8). Distinct patterns from the Pleistocene in relation to the Holocene are given for the subcompositions (aragonite, quartz, kaolinite) in Figure 8D, or for the subcomposition (aragonite, quartz, albite) in Figure 8C. The pattern becomes clearer by amalgamation of the carbonates and clay minerals (Fig. 8A).


Figure 8: Selected ternary diagrams

A sequential binary partition for the mineral data set is proposed in Table 5. The associated balances offer the following interpretation approaches: (b1) importance of aluminosilicates, against carbonate and quartz; (b2) importance of primary carbonate production (aragonite); (b3) origin of calcite (detrital vs. chemically precipitated); (b4) clays of the type $2: 1$ against feldspar and $1: 1$ clays; (b5) K bearing versus Fe and Mg bearing clays; (b6) importance of weathering of albite into kaolinite.

Table 5: possible binary partition for the major mineral composition (quartz, calcite, aragonite, muscovite/illite, chlinoclore, kaolinite, albite).

| balance | Quartz | Calcite | Aragonite | Musc./Illite | Chlinoclore | Kaolinite | Albite |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $\mathbf{b 1}$ | +1 | +1 | +1 | -1 | -1 | -1 | -1 |
| $\mathbf{b 2}$ | -1 | -1 | +1 | 0 | 0 | 0 | 0 |
| b3 | -1 | +1 | 0 | 0 | 0 | 0 | 0 |
| b4 | 0 | -1 | 0 | +1 | +1 | -1 | -1 |
| b5 | 0 | 0 | 0 | +1 | -1 | 0 | 0 |
| b6 | 0 | 0 | 0 | 0 | 0 | +1 | -1 |



Figure 9: (A) balance dendogram for the major mineral assemblage; (B) zoom into balances b4 and b6; (C) zoom into balance b3.

Within the balance dendogram in Figure 9, again, highest variance lies between the carbonates and the detrital input. For the balances relating carbonates and detrital, sediment unit IV holds the highest variances, because this sedimentary unit retains a highly detrital character, and reduced Ca accumulation (Fig. 1, and Fig. 9 b2), expressed by nearly absent aragonite (Fig. 8B, 8C, and 8D). Balance 3 shows a very small variance of calcite related to the detrital quartz input, leading to the inference that calcite has a detrital character. Although the high variance from calcite to the siliciclastics, as well as to the primary produced aragonite (Fig.7) indicates that calcite accumulation and preservation in Iznik sediments is manifold. For instance in addition to the possible detrital or primary origins, the shells of ostracodes are
made of biogenically precipitated calcite. Figure 9B zooms into the ratio between kaolinite and albite, which increases for the Holocene (sediment units I and II). The chemical weathering of the feldspar albite originates kaolinite, and dissolved Na as by products (Berner, 1971, and 1980). The observations from Figure 9 balance 6, as well as the patterns from ternary subcompositions ( $\mathrm{Ca}, \mathrm{Na}, \mathrm{Ti}$ ) in Figure 5A and (carbonates, quartz+albite, clays) in Figure 8A strongly suggest increased chemical weathering for sediment unit I, implying a increased atmospheric moisture supply for the Marmara region, during the middle Holocene.

## 4 Conclusions

Based on absolute concentrations of elements, on average silica is more abundant than calcium in Iznik sediments, however data also shows that a great amount of geochemical variability is retained by calcium, and mineralogical variability by aragonite. Hence, Lake Iznik is unique for its carbonate accumulation in the form of aragonite, which is favored by distinctive water chemistry of the lake.

Overall, the clr-biplots of geochemical data, show three main groups of elements: (1) those that present low variance to calcium, (2) those that present low variance to the organic carbon, and (3) siliciclastic elements. Nevertheless, some elements present a slight higher variance in relation to the siliciclastic phase, for example: manganese - which is subject to early diagenetical reactive transport due to changes in the redox conditions; and sodium - which is impoverished in the Holocene sediments.

A distinct pattern observed for the geochemical ternary subcomposition $(\mathrm{Ca}, \mathrm{Na}, \mathrm{Ti})$ in Figure 5 A , is observed also for the mineral ternary subcomposition (carbonates, quartz+albite, clays) in Figure 8A, suggesting that the impoverishment in sodium accumulation during the middle Holocene results from the enhanced chemical weathering of albite in the basin.

The mineral clr -biplot displays analogies to the geochemical biplot. The siliciclastic and carbonate groups identified in the geochemical clr-biplot refer to the detrital minerals, and aragonite in the mineral clrbiplot, respectively. Calcite clearly has an additional and distinct source in relation to the siliciclastic elements, as well as to the primary carbonate aragonite. In fact, this mineral also builds shells of ostracodes.

Overall, this study shows that a parallel exploratory data analysis of two different compositional data sets (in this case geochemical and mineralogical) from the same geological matrix can give meaningful insights about the genesis of such geological material. The development of appropriate compositional methods to explore patterns or relationships in between such compositional data sets is highly desirable.

## References

Aitchison, J. 2003a. The Statistical Analysis of Compositional Data. 2nd ed: The Blackburn Press.
Aitchison, J. 2003b. A Concise Guide to Compositional Data Analysis. Lecture Notes. Available online at CoDaWeb.

Aitchison, J., and Greenacre, M. 2002. Biplots of compositional data. Journal of the Royal Statistical Society Series C-Applied Statistics 51:375-392.

Berner, R. A. 1971. Principles of Chemical Sedimentology, International Series in the Earth and Planetary Sciences: McGraw-Hill Book Company.

Berner, R. A. 1980. Early Diagenesis - A Theoretical Approach. New Jersey: Princeton University Press.
Comas-Cufí M, Thió-Henestrosa S. CoDaPack 2.0: a stand-alone, multi-platform compositional software. In: Egozcue JJ, Tolosana-Delgado R, Ortego MI, eds. CoDaWork'11: 4th International Workshop on Compositional Data Analysis. Sant Feliu de Guíxols; 2011.

Egozcue, J. J., and Pawlowsky-Glahn, V. 2006. Simplicial geometry for compositional data. Geological Society, London, Special Publications 264 (1):145-159.

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

Moore, D. M., and Reynolds Jr., R. C. 1997. X-Ray diffraction and the identification and analysis of clay minerals. 2nd Edition ed. Oxford: Oxford University Press

Pawlowsky-Glahn, V., Egozcue J. J., and Tolosana-Delgado, R. Lecture Notes on Compositional Data Analysis. pp. 96. Girona, Barcelona: Univeristy of Girona, Technical University of Catalonia, 2011.

Roeser, P. (2014). Paleolimnology of Lake Iznik, PhD Thesis, pp. 201., University of Bonn

Roeser, P. A., Franz, S. O., Litt, T., Ülgen U. B., Hilgers, A., Wulf, S., Wennrich, V., Akçer Ön, S., Viehberg, F. A., Çağatay, N. M., and Melles, M. 2012. Lithostratigraphic and geochronological framework for the paleoenvironmental reconstruction of the last $\sim 36 \mathrm{ka} \mathrm{cal} \mathrm{BP} \mathrm{from} \mathrm{a} \mathrm{sediment}$ record from Lake Iznik (NW Turkey). Quaternary International 274 (0):73-87

Tolosana-Delgado, R. 2012. Uses and misuses of compositional data in sedimentology. Sedimentary Geology 280 (0):60-79.

# Joint compositional calibration: a geochronological example 

R. Tolosana-Delgado ${ }^{1}$, K. G. van den Boogaart ${ }^{1}$, E. Fiserova ${ }^{2}$, K. Hron ${ }^{2}$, and I. Dunkl ${ }^{3}$<br>${ }^{1}$ Helmholtz Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resource Technology, Freiberg, Germany r.tolosana@hzdr.de<br>${ }^{2}$ Palacky University, Dept. Mathematical Analysis and Applications of Mathematics, Olomouc, Czech Republic<br>${ }^{3}$ Georg August University, Dept. Sedimentology and Environmental Geology, Goettingen, Germany


#### Abstract

This contribution explores several issues arising in the measurement of a (geo)chemical composition, specially in the case that the quantities of interest are linear functions of (log)-ratios. This is the case of multi-isotopic geochronological analyses, where an age of a geological material is inferred as a function of several logratios in the isotopic system (Hg202, Pb204, Pb206, Pb207, Pb208, Th232, U235, U238), by using several ratios (e.g., Pb208/Pb206, Pb206/Pb204, Hg202/Pb204, U238/U235), quantities that are scaling invariant and subcompositionally coherent, but which in general cannot be estimated without taking into account additive noise effects incompatible with a purely compositional approach. The proposed ways to a solution heavily build upon the multiPoisson distribution, higlighting the counting nature of the readings delivered by these measuring machines.


## 1 Introduction

Elemental abundances are obtained with LA-ICP-MS (Laser Ablation Inductively Coupled Plasma Mass Spectrometry): the matter to analyse is ablated with a laser and the resulting gas is introduced in a plasma, where its atoms part with their most loosely bound electron to form a $(+1)$-charged ion. These are then separated by a magnetic field, on the basis of their mass-charge ratio, each colliding with a detector. The result is a vector of counts (collisions per unit of time) for several possible masses, which must then somehow be related to concentrations or relative abundances of the several elements or isotopes present, usually in a proportional way.

Thus, the machine is often used alternatively with analytes of unknown composition (samples) or of known composition (standards), in order to derive the proportionality factor between signal and composition. It is worth noting that the detectors receive counts even when no analyte is present (background or blank, due to e.g. rest samples, other gases in the plasma, etc.), and that the signal received while analysing standards shows systematic drifts at several time scales (along the day, during a measurement, etc).

Thus, the desired ratios of abundances must be estimated from data on counts, taking into account the noise in the signal and a (potenially non-constant) proportionality between ratios of signals and ratios of abundances along time. This setting is often attacked by calibrating each isotope abundance separately, subtracting an additive noise and assuming a proportionality function on time interpolated between standards. Several interpolation methods have been reported to be used, though mostly they are piecewise linear interpolations or polynomial fits.

None of the existing methods considers in any sense the possible compositional nature of the problem. This contribution explores the reasons behind that and presents several ways of taking that into account. This work builds upon materials presented at the previous edition of CoDaWork (2013).

## 2 Basics of Laser Ablation Inductively Coupled Plasma Mass Spectrometry (LA-ICP-MS)

### 2.1 Recall of Poisson distribution properties

A random variable $X$ is said to follow a Poisson distribution with intensity parameter $\lambda$, denoted as $X \sim \mathcal{P} o(\lambda)$ if and only if its probability density function is

$$
f_{X}(x)=\frac{\lambda^{x}}{x!} e^{-\lambda}, \quad x=0,1,2,3, \ldots
$$

The expected value and population variance of $X$ are $\mathrm{E}[X]=\operatorname{Var}[X]=\lambda$, hence this parameter is often called expected number of counts. Note that, in spite of this identification, $X$ is an integer, while the parameter and these statistics are real positive values $\lambda \in \mathbb{R}_{+}$. Its dispersion coefficient defined as $D=\operatorname{Var}[X] / \mathrm{E}[X]=1$ always. Note that this is not the coefficient of variation.

If $\lambda$ is large, the Poisson distribution can be excellently approximated by a normal distribution with mean and variance both equal to $\lambda$.

The sum of two independent Poisson distributed variates $X_{1} \sim \mathcal{P} o\left(\lambda_{1}\right)$ and $X_{2} \sim \mathcal{P} o\left(\lambda_{2}\right)$ follows as well a Poisson distribution $Y=X_{1}+X_{2} \sim \mathcal{P} o\left(\lambda_{1}+\lambda_{2}\right)$. The difference $Z=X_{1}-X_{2}$ follows a Skellam distribution (Skellam, 1946), but if $\lambda_{1} \gg \lambda_{2}$, then approximately $Y=X_{1}-X_{2} \sim$ $\mathcal{P} o\left(\lambda_{1}-\lambda_{2}\right)$.

A vector of $D$ Poisson variates $X_{i} \sim \mathcal{P} o\left(\lambda_{i}\right)$ follows a multi-Poisson distribution with parameter vector $\boldsymbol{\lambda}=\left[\lambda_{1}, \lambda_{2}, \ldots, \lambda_{D}\right]$. This is a vector of non-negative integer components characterized by the following conditional construction:

1. the sum of these components $X^{T}=\sum_{i}^{D} X_{i} \sim \mathcal{P} o\left(\lambda^{T}=\sum_{i}^{D} \lambda_{i}\right)$ gives a total number of counts; note that this holds because the several components are independent by the nature of the Poisson process;
2. conditional on a drawn total number of counts $x^{T}$, the number of counts on each category follows a multinomial distribution with parameters $\mathbf{p}=\mathcal{C}[\boldsymbol{\lambda}]$ and $n=x^{T}$.

### 2.2 Technical procedure

Laser ablation inductively coupled plasma mass spectrometer (LA-ICP-MS) is an in situ analytical instrument that is used for determining the elemental and isotopic composition of micrometer-sized areas of solid materials such as minerals, glasses, metal alloys, bones, teeth, calcareous shell and wood, and fluids trapped as inclusions within solids. The analytes are polished and cleaned before introduction into the laser cell. Typically an excimer UV laser ( 193 nm ) is used for ablation, the generated aerosol from the ca. 20 to $100 \mu \mathrm{~m}$ diameter laser spot is transported by a Helion-Argon gas mixture into the ICP-MS. Here all atoms are excited to plasma state by removal of one electron of each atom. The resulting +1 ions are accelerated in a magnetic field, which makes them take different, well-defined, trajectories according to their mass-to-charge ratio: i.e. given that the charge is always +1 , according to their atomic mass.

A sensor placed at the end of the trajectory of a given mass counts the number of impacts of atoms of that mass during a certain time period (dwell time). Then the sensor moves to a second position, corresponding to a second mass, and counts impacts there. And so on, until the complete set of masses has been visited. Afterwards the sensor comes back to the starting position and counts again, thus starting a new time slice. Several of such time slices occur during a second, producing a multivariate reading. In spite of the sequential character of the measurements within a time slice, these are considered simultaneous.

This reading procedure is applied in three different situations:
blank the readings obtained with no material,
standard the readings obtained with an analyte of known composition,
sample (s.s.) with an analyte of unknown composition.

Given that the counts are registered for each atomic mass during a fraction of a second (the dwell time, which might be different for each element), the resulting total counts per time window are linearly upscaled to counts per second.

### 2.3 Conventional data analysis approach

Typically, each measuring interval $T_{k}$ contains a period of blank readings and a period of sample or standard readings, with more or less sharp transitions between them. The first step is the identification of time windows of blank $\left[t_{A}, t_{B}\right] \subset T_{k}$ and of measurement $\left[t_{M}, t_{N}\right] \subset T_{k}$ (see Fig. 3 later on, for an example), a task that is often manually done by the lab analyst guided by some homogeneity statistics.

The second step is the characterization of the blank. This is usually done for each measuring period and for each istope $i$ separately. All readings in the blank window, denoted $\left\{x_{i}\left(t_{A}\right), x_{i}\left(t_{A}+\right.\right.$ $\left.1), \ldots, x_{i}\left(t_{B}\right)\right\}$, are considered as independent realizations of a Poisson distribution of unkwown parameter $\lambda_{i}$. With the standard assumptions of statistics of Poisson variates, this parameter can be estimated as the mean readings of that variable in the blank window, denoted as $b_{k i}$. Some labs implement a quality control assessment on the dispersion coefficient

$$
\hat{D}=\frac{\operatorname{Var}\left[x_{i}\left(t_{A}\right), x_{i}\left(t_{A}+1\right), \ldots, x_{i}\left(t_{B}\right)\right]}{\mathrm{E}\left[x_{i}\left(t_{A}\right), x_{i}\left(t_{A}+1\right), \ldots, x_{i}\left(t_{B}\right)\right]}=\frac{s_{b i}^{2}\left(T_{k}\right)}{b_{k i}}
$$

with heuristic rules that suggest a too-strong non-Poissonal regime if the hypothesis $\hat{D}=1$ is not acceptable. In this case, typically the analysts reconsider their choice of blank window.

The third step is the definition of the expected readings for each sample or standard measuring window. This is sometimes applied to the absolute readings (in case of LA-ICP-MS measurements of concentrations), but in geochronological studies the following ratios of interest are rather used: $\mathrm{Pb}^{206} / \mathrm{U}^{234}, \mathrm{~Pb}^{207} / \mathrm{U}^{235}, \mathrm{~Pb}^{207} / \mathrm{Pb}^{206}$ and eventually $\mathrm{Pb}^{208} / \mathrm{Th}^{232}$. A first approach would be to neglect fractionation effects on those ratios and apply the same procedure of the blank at the set of readings $\left\{\Delta x_{i}\left(t_{M}\right), \Delta x_{i}\left(t_{M}+1\right), \ldots, \Delta x_{i}\left(t_{N}\right)\right\}$ vs. $\left\{\Delta x_{j}\left(t_{M}\right), \Delta x_{j}\left(t_{M}+1\right), \ldots, \Delta x_{j}\left(t_{N}\right)\right\}$. Note that these values are obtained subtracting to the read counts the blank levels $b_{k i}$ and $b_{k j}$. Some labs work with the arithmetic mean of the ratios $\Delta x_{i}\left(t_{m}\right) / \Delta x_{j}\left(t_{m}\right)$, while other work with the ratios of the count means $\overline{\Delta x} i / \overline{\Delta x}_{j}$ within the measuring window. More elaborate approaches consider the fractionation trend as a line or as a curve, and attempt several ways of extracting a representative average ratio. For instance, an option is to fit a linear regression trend to the fractionation drift and extrapolate it to the time moment when the analyte plasma arrived at the chamber. Whichever method is used, for each measuring interval one has a value blank value and a measurement value for each quantity of interest (ratio or concentration) $y_{k l}$ where $k$ is the index of the measuring interval and $l$ is the quantity of interest.

The fourth step is to study the several measurements $\left\{y_{k l}\right\}$ available for the standards, which due to their homogeneity, should be "equal", i.e ideally realizations of the same random variable. If this can be assumed, then the average of all standard readings $y_{l}^{s t d}$ is compared with the known
nominal value $\mu_{l}^{s t d}$, and all measurements for unknown analytes are upscaled conveniently as

$$
y_{k l}=\frac{\mu_{l}^{s t d}}{y_{l}^{s t d}} y_{k l} .
$$

If the several readings of the standard show too obvious systematic drifts, then it is common to assume some functional model for this drift, fit it to the available measurements of the standard, predict its value $y_{l}^{s t d}\left(T_{k}\right)$ for a measuring interval $T_{k}$ and then upscale the value at that interval accordingly

$$
y_{k l}=\frac{\mu_{l}^{s t d}}{y_{l}^{s t d}\left(T_{k}\right)} y_{k l} .
$$

Some effort in these data processing steps is devoted to evaluating the "measurement error" (statistical error, uncertainty). Denoting by $\bar{x}_{i}, s_{x i}^{2}$ and $b_{i}, s_{b i}^{2}$ the means and variances of the signal (no blank correction) and of the background (dropping the dependence on $T_{k}$ for simplicity) of the isotope $i$, the variance of the corrected signal is

$$
s_{c i}^{2}=s_{b i}^{2}+s_{x i}^{2},
$$

under the hypothesis that the blank and the corrected signal are considered as independent Poisson distributions which sum gives the uncorrected signal. These concepts and statistics allow as well the definition of detection limit. The detection limit is defined as that level of signal which cannot be distinguished from the blank. It is customary to take the detection limit of each isotope counts as $D L_{i}=b_{i}+3 \cdot s_{b i}$.

## 3 Compositional calibration models for ICP-MS

### 3.1 Amount-scalar upscaling model

Following the ideas of the practitioners, we assume the blank vector of $D$ components to be a random realization of a multi-Poisson distribution, i.e. at each time slice

$$
\mathbf{X}\left(t_{k}\right) \sim \mathcal{P} o\left(\omega_{0} \boldsymbol{\lambda}_{b}\right), \quad t_{k} \in\left[t_{A}, t_{B}\right],
$$

where $\boldsymbol{\lambda}_{b}=\left[\lambda_{b 1}, \lambda_{b 2}, \ldots, \lambda_{b D}\right]$ is the expected counts per second of the blank and $\omega_{0}$ is the length of the dwell time in seconds. The corrected signal at a given time slice $t_{k}$ while analysing a sample of amounts $\mathbf{z}$ is assumed to be

$$
\begin{equation*}
\mathbf{X}\left(t_{k}\right) \sim \mathcal{P} o\left(\omega_{0} \lambda\left(t_{k}\right) \mathbf{z}\right), \quad t_{k} \in\left[t_{M}, t_{N}\right], \tag{1}
\end{equation*}
$$

where $\lambda\left(t_{k}\right)$ is an upscaling factor that describes how many counts per second are expected for a gram of the sample. The quantity $\lambda\left(t_{k}\right)$ captures the time-varying sensitivity of the machine. Note that Eq. (1) implies a non-compositional nature of the process, as the intensity of the Poisson process is not scaling invariant. The quantity $\left(\omega_{0} \lambda\left(t_{k}\right) \mathbf{z}\right)$ is actually rather an object of $\mathbb{R}_{+}^{D}$, the multivariate positive real space.

Under the parsimonious assumption of independent signal and blank, the signal read is

$$
\mathbf{X}\left(t_{k}\right) \sim \mathcal{P} o\left(\omega_{0}\left(\lambda\left(t_{k}\right) \mathbf{z}+\boldsymbol{\lambda}_{b}\right)\right) .
$$

This equation is actually also valid for the blank, because in that case $\mathbf{z}=\mathbf{0}$. Again, the noncompositional nature of this equation must be highlighted, in this case derived from the additivity properties of the Poisson distribution.

Note as well that the counts $\mathbf{X}\left(t_{k}\right)$ follow a multinomial distribution with parameter vector $\mathcal{C}\left[\lambda\left(t_{k}\right) \mathbf{z}+\boldsymbol{\lambda}_{b}\right]$, and $n$ the total number of counts, with $n \sim \mathcal{P}_{O}\left(\omega_{0} \mathbf{1}^{\prime} \cdot\left(\lambda\left(t_{k}\right) \mathbf{z}+\boldsymbol{\lambda}_{b}\right)\right)$. This model assumes that the sensitivity $\lambda\left(t_{k}\right)$, though time-varying, is the same for all elements.

### 3.2 Amount-perturbation upscaling model

A first modification of the preceding model is to assume a different sensitivity parameter for each element, but no interaction between the several elements, i.e. that:

- one gram of pure element $i$ produces $\lambda_{i}\left(t_{k}\right)$ expected counts; and
- the counts of two grams of the same element are simply added together, i.e. that mass is additive.

Thus a mass $z_{i}$ of element $i$ produces $z_{i} \lambda_{i}\left(t_{k}\right)$ expected counts, and a vector of amounts $\mathbf{z}$ produces total expected counts $\lambda\left(t_{k}\right)=\mathbf{z}^{\prime} \cdot \boldsymbol{\lambda}\left(t_{k}\right)=\sum_{i} z_{i} \cdot \lambda_{i}\left(t_{k}\right)$. The actual number of total counts obtained by the machine $n$ is then distributed as $n \sim \mathcal{P} o\left(\omega_{0}\left(\lambda\left(t_{k}\right)+\mathbf{1}^{\prime} \cdot \boldsymbol{\lambda}_{b}\right)\right)$, and these are split in the $D$ elements following a multinomial with parameters $n$ and $\mathcal{C}\left[\mathbf{z} \oplus \boldsymbol{\lambda}\left(t_{k}\right)+\boldsymbol{\lambda}_{b}\right]$. In this expression, $\oplus$ denotes the perturbation in $\mathbb{R}_{+}^{D}$, i.e. without closure (Pawlowsky-Glahn and Egozcue, 2001; van den Boogaart and Tolosana-Delgado, 2008).

### 3.3 Amount-matrix upscaling-interaction model

Finally, the last option implies interaction between the counts of the several elements, i.e. that a vector of masses $\mathbf{z}$ produces a vector of expected counts $\Lambda(\mathbf{z})$ with $\Lambda(\cdot)$ an application between $\mathbb{R}_{+}^{D}$ and $\mathbb{R}_{+}^{D}$. Given the additive character of both masses and counts, it is reasonable to assume that $\Lambda(\mathbf{z})$ can be considered a linear form $\boldsymbol{\Lambda} \cdot \mathbf{z}$, i.e. $\boldsymbol{\Lambda}=\left[\lambda_{i j}\right]$ is a matrix of $(D, D)$-elements, where $\lambda_{i j}$ is the contribution of component $j$ to the counts of component $i$.

In general, all elements $\lambda_{i j}\left(t_{k}\right)$ represent time-dependent sensitivities, which would be quite difficult to estimate. However, it is reasonable to assume that the relative sensibility interaction coefficients $\lambda_{i j}^{*}=\lambda_{i j}\left(t_{k}\right) / \lambda_{j j}\left(t_{k}\right)$ are quasi constant, meaning that the counts generated by a mass $z_{j}$ are split among the $D$ components in fractions which do not change with time. In this case, a vector of amounts $\mathbf{z}$ produces a vector of expected counts $\boldsymbol{\Lambda}^{*} \cdot\left(\mathbf{z} \oplus \boldsymbol{\lambda}\left(t_{k}\right)\right)$, with a constant interaction matrix $\boldsymbol{\Lambda}^{*}=\left[\lambda_{i j}^{*}\right]$, dominantly diagonal with elements $\lambda_{i j}=1$, and a time-dependent sensitivity vector $\boldsymbol{\lambda}\left(t_{k}\right)$. Adding the blank, we obtain a model where the total number of counts $n$ is distributed
as a Poisson with intensity $\omega_{0} \mathbf{1}^{\prime} \cdot\left(\boldsymbol{\Lambda}^{*} \cdot\left(\mathbf{z} \oplus \boldsymbol{\lambda}\left(t_{k}\right)\right)+\boldsymbol{\lambda}_{b}\right)$, and these counts are split among the $D$ components following a multinomial with parameters $n$ and $\mathcal{C}\left[\boldsymbol{\Lambda}^{*} \cdot\left(\mathbf{z} \oplus \boldsymbol{\lambda}\left(t_{k}\right)\right)+\boldsymbol{\lambda}_{b}\right]$.

Note that these expressions do not entirely correspond to any geometry of those summarized by Pawlowsky-Glahn and Egozcue (2001) or van den Boogaart and Tolosana-Delgado (2008) for the simplex, the positive real space $\mathbb{R}_{+}^{D}$ or the real space $\mathbb{R}^{D}$ : perturbation $\oplus$ relates in this case to the Abelian group operation on $\mathbb{R}_{+}^{D}$; addition + and the linear form $\boldsymbol{\Lambda}^{*}$ act on the classical Euclidean geometry of $\mathbb{R}^{D}$, and the closure $\mathcal{C}[\cdot]$ projects the combination of signal and background onto the simplex. This is a typical construction in multi-Poisson processes.

### 3.4 Models ignoring the blank or with multiplicative effects

From the point of view of (log-ratio based) compositional data analysis, all models derived before are particularly complicated by the presence of the additive blank. If this could be ignored (e.g. because it is very small with regard to the signal), and the sum of the components of $\mathbf{z}$ equals one (i.e. a whole composition is analysed), then the following models are derived:

- scalar upscaling: $\left[\mathbf{X}\left(t_{k}\right) \mid n\right] \sim \mathcal{M} u(\mathbf{z} ; n)$ and $n \sim \mathcal{P} o\left(\omega_{0} \lambda\left(t_{k}\right)\right)$;
- perturbation: $\left[\mathbf{X}\left(t_{k}\right) \mid n\right] \sim \mathcal{M} u\left(\mathbf{z} \oplus \boldsymbol{\lambda}\left(t_{k}\right) ; n\right)$ and $n \sim \mathcal{P} o\left(\omega_{0} \sum_{i}^{D} \lambda_{i}\left(t_{k}\right)\right)$, in this case with $\oplus$ the perturbation on the simplex;
- interaction-perturbation: $\left[\mathbf{X}\left(t_{k}\right) \mid n\right] \sim \mathcal{M} u\left(\mathcal{C}\left[\boldsymbol{\Lambda}^{*} \cdot\left(\mathbf{z} \oplus \boldsymbol{\lambda}\left(t_{k}\right)\right)\right] ; n\right)$ and $n \sim \mathcal{P} o\left(\omega_{0} \mathbf{1}^{\prime} \cdot\left(\boldsymbol{\Lambda}^{*}\right.\right.$. $\left.\left(\mathbf{z} \oplus \boldsymbol{\lambda}\left(t_{k}\right)\right)\right)$ ), in this case with $\oplus$ the perturbation on $\mathbb{R}_{+}^{D}$.

This last model is still a mixture of additive and multiplicative geometries. The following models are purely compositional alternatives, using in the compositional part only operations on the simplex

- scalar upscaling: $\left[\mathbf{X}\left(t_{k}\right) \mid n\right] \sim \mathcal{M} u\left(\mathbf{z} \oplus \boldsymbol{\lambda}_{b} ; n\right)$ and $n \sim \mathcal{P} o\left(\omega_{0} \lambda\left(t_{k}\right) \sum_{i}^{D} \lambda_{b i}\right)$;
- perturbation: $\left[\mathbf{X}\left(t_{k}\right) \mid n\right] \sim \mathcal{M} u\left(\mathbf{z} \oplus \boldsymbol{\lambda}\left(t_{k}\right) \oplus \boldsymbol{\lambda}_{b} ; n\right)$ and $n \sim \mathcal{P} o\left(\omega_{0} \sum_{i}^{D} \lambda_{b i} \lambda_{i}\left(t_{k}\right)\right)$;
- interaction-perturbation: $\left[\mathbf{X}\left(t_{k}\right) \mid n\right] \sim \mathcal{M} u\left(\boldsymbol{\Lambda}^{*} \square\left(\mathbf{z} \oplus \boldsymbol{\lambda}\left(t_{k}\right)\right) \oplus \boldsymbol{\lambda}_{b} ; n\right)$ and $n \sim \mathcal{P} o\left(\omega_{0} \mathbf{1}^{\prime} \cdot\left(\boldsymbol{\Lambda}^{*} \square\right.\right.$ $\left.\left.\left(\mathbf{z} \oplus \boldsymbol{\lambda}\left(t_{k}\right)\right) \oplus \boldsymbol{\lambda}_{b}\right)\right)$, in this case with $\oplus$ the perturbation on the simplex.

In these expressions, we have used the notation $\square$ after Pawlowsky-Glahn, Egozcue and TolosanaDelgado (2015, Chap. 4) to denote a simplicial endomorphism operation, i.e. one such that once expressed in any basis of the simplex becomes a simple matrix-vector product. Note that these purely compositional models imply, among other effects, that the noise induced by the blank upscales with the signal, i.e. larger signal should show more blank variability.

### 3.5 Minor modifications of the preceding models

As mentioned, sensitivities can be equal for all elements or have a different value for each element, be an endomorphism is some geometry (hence a matrix) or even be a general functional; and
they can be constant on time or show a trend, which we will parameterize through a vector $\boldsymbol{\theta}$. Common cases for trends are polynomial trends, or piecewise linear trends. For the sake of the illustration later on, we will assume a linear trend, e.g. for an amount-perturbation model, $\boldsymbol{\Lambda}\left(\mathbf{z}\left(t_{i}\right), \boldsymbol{\lambda}_{b}, \boldsymbol{\theta} ; t_{i}\right)=\mathbf{z} \oplus \boldsymbol{\theta}_{0} \oplus t \odot \boldsymbol{\theta}_{t}+\boldsymbol{\lambda}_{b}$, i.e. with some initial sensitivities $\boldsymbol{\theta}_{0}$ and their rates of change along time $\boldsymbol{\theta}_{t}$, different for each isotope and without interaction.

In any case, we finally have distributions for the number of counts on each element class that belong to the multi-Poisson family, with an intensity vector model $\omega_{0} \boldsymbol{\Lambda}\left(\mathbf{z}, \boldsymbol{\lambda}_{b}, \boldsymbol{\theta} ; t_{k}\right)$ capturing the relationship between the expected partial counts and the composition of the analyte. Hence, we can always consider that the total number of counts $n\left(t_{k}\right)$ follows a Poisson distribution with $\lambda_{k}^{T}=\omega_{0} \mathbf{1}^{\prime} \cdot \boldsymbol{\Lambda}\left(\mathbf{z}, \boldsymbol{\lambda}_{b}, \boldsymbol{\theta} ; t_{k}\right)$; and, conditional on that total, the vector of counts for each element follows a multinomial distribution with probability parameter vector $\mathbf{p}_{k}=\mathcal{C}\left[\boldsymbol{\Lambda}\left(\mathbf{z}, \boldsymbol{\lambda}_{b}, \boldsymbol{\theta} ; t_{k}\right)\right]$.

A relevant minor modification for the example we will present later consists of the case that the dwell times are not equal for all isotopes. If we denote the vector of dwell times as $\boldsymbol{\omega}_{0} \in \mathbb{R}_{+}^{D}$, then it is immediate to show that the number of counts on each element class belong to the multiPoisson family, with a perturbed intensity vector model $\boldsymbol{\omega}_{0} \oplus \boldsymbol{\Lambda}\left(\mathbf{z}, \boldsymbol{\lambda}_{b}, \boldsymbol{\theta} ; t_{k}\right)$, with perturbation on $\mathbb{R}_{+}^{D}$. Again, the total number of counts $n\left(t_{k}\right)$ will follow a Poisson distribution, albeit with total expected counts $\lambda_{k}^{T}=\boldsymbol{\omega}_{0}^{\prime} \cdot \boldsymbol{\Lambda}\left(\mathbf{z}, \boldsymbol{\lambda}_{b}, \boldsymbol{\theta} ; t_{k}\right)$; and the vector of counts for each element conditionally follows a multinomial distribution with probability parameter vector $\mathbf{p}_{k}=\mathcal{C}\left[\boldsymbol{\omega}_{0} \oplus \boldsymbol{\Lambda}\left(\mathbf{z}, \boldsymbol{\lambda}_{b}, \boldsymbol{\theta} ; t_{k}\right)\right]$.

## 4 Methods

### 4.1 Notation and common asumptions

Let us assume one particular model $\boldsymbol{\Lambda}\left(\mathbf{z}\left(t_{i}\right), \boldsymbol{\lambda}_{b}, \boldsymbol{\theta} ; t_{i}\right)$ from those mentioned before. A set of readings of count vectors $\left\{\mathbf{x}\left(t_{i}\right), t_{i} \in T\right\}$ is available, obtained along a session $T$ split in $K$ intervals $T_{1}, T_{2}, \ldots, T_{K}$. Each interval contains two non-overlapping windows, the blank window $B_{k} \subset T_{k}$ and the measurement window $M_{k} \subset T_{k}$. All readings during the blank window are obtained with no material being analysed, i.e. $\mathbf{z}\left(t_{i}\right)=\mathbf{0}, \forall t_{i} \in B=\bigcup_{k=1}^{K} B_{k}$. During the measuring window $M_{k}$ an analyte of different composition was analysed, i.e. $\mathbf{z}\left(t_{i}\right)=\mathbf{z}_{k}$. Some of these compositions $\left\{\mathbf{z}_{1}, \mathbf{z}_{2}, \ldots, \mathbf{z}_{K}\right\}$ are known (those corresponding to standards) and some other are unknown and the actual target of this problem (the samples). Let the set of indices $\mathcal{K}=\{1,2, \ldots, K\}$ be partitioned in two disjoint subsets $\mathcal{K}_{s}$ (corresponding to the time intervals when a standard was analysed) and $\mathcal{K}_{m}$ (corresponding to the intervals when a sample of unknown composition was analysed). The goal is thus to estimate all $\mathbf{z}_{k}$ for $k \in \mathcal{K}_{m}$, given the set of all observations $\left\{\mathbf{x}\left(t_{i}\right), t_{i} \in T\right\}$ and the composition of the standards $\mathbf{z}_{k}$ for $k \in \mathcal{K}_{s}$. The set of data can also be split in measurements corresponding to all blank periods and measurement of standards $\mathbf{X}_{s}$, the calibration set $\mathbf{X}^{s}=\left\{\mathbf{x}\left(t_{i}\right), t_{i} \in B \cup \bigcup_{k \in \mathcal{K}_{s}} M_{k}\right\}$; and data correspond to measurement windows of samples of unknown composition, the prediction set $\mathbf{X}^{m}=\left\{\mathbf{x}\left(t_{i}\right), t_{i} \bigcup_{k \in \mathcal{K}_{m}} M_{k}\right\}$.

Given this setting, we will assume a multi-Poisson distribution for the data $\left\{\mathbf{x}\left(t_{i}\right), t_{i} \in T\right\}$, with intensity model $\boldsymbol{\Lambda}\left(\mathbf{z}\left(t_{i}\right), \boldsymbol{\lambda}_{b}, \boldsymbol{\theta} ; t_{i}\right)$. This will allow us eventually to study this system in the two contributions, i.e. with total counts Poisson distributed, and these counts then as Multinomial
distributed. We will as well assume two different sensitivity models:

- an amount perturbation-scaling model,
- some form of a amount-matrix upscaling-interaction model (be it an an additive or a multiplicative geometry).


### 4.2 Generalized linear model (GLM)

To apply the formalism of generalized linear models (Nelder and Wedderburn, 1972) we need to further assume that the composition of the standards is perfectly known and homogeneous. In this case we can consider the problem divided in two steps:

1. Calibration phase. In this phase we consider only the data available from all blank windows and the measurement windows of the standard.
2. Prediction phase. In this phase the model is used to predict the composition of the unknown samples.

### 4.2.1 Amount-perturbation model with time drift

This case is the easiest to understand, as the lack of any form of interaction allows to estimate each component separately. In this case we have

$$
X_{j}^{s}\left(t_{i}\right) \sim \mathcal{P} o\left(\omega_{0 j}\left[\left(\theta_{0 j}+\theta_{j} t_{i}\right) z_{j}\left(t_{i}\right)+\lambda_{b j}\right]\right) \sim \mathcal{P} o\left(a_{j}+b_{j} z_{j}\left(t_{i}\right)+c_{j} z_{j}\left(t_{i}\right) t_{i}\right)
$$

In the calibration phase, given that $z_{j}\left(t_{i}\right)$ and $t_{i}$ are known everywhere, this model can be estimated with a generalized linear model of the Poisson family with identity link, a non-canonical choice. In this step $a_{j}=\omega_{0 j} \lambda_{b j}, b_{j}=\omega_{0 j} \theta_{0 j}$ and $c_{j}=\omega_{0 j} \theta_{j}$ are the parameters to be estimated. In the prediction phase, with estimates $\tilde{a}_{j}, \tilde{b}_{j}, \tilde{c}_{j}$ set, the unknown is $z_{j}\left(t_{i}\right)$, i.e. the model becomes again a generalized linear model

$$
X_{j}^{m}\left(t_{i}\right) \sim \mathcal{P} o\left(\tilde{a}_{j}+\left[\tilde{b}_{j}+\tilde{c}_{j} t_{i}\right] z_{j}\left(t_{i}\right)\right)
$$

where $\tilde{a}_{j}$ is an offset and the predictor variable $\left[\tilde{b}_{j}+\tilde{c}_{j} t_{i}\right]$ is known everywhere. This model has no intercept allowed. These models can be estimated in both steps with the GLM maximization likelihood procedures.

Note that the canonical choice of the Poisson family (the logarithmic link) would give rise to the multiplicative perturbation-scaling model, but it is not compatible with the physics of the phenomenon, dominated by the additive character of mass.

### 4.2.2 Amount-matrix upscaling-interaction model with time drift

In this case, given the interaction between components it is not possible to consider them totally independently any more. Considering a linear drift $\boldsymbol{\Lambda}\left(t_{i}\right)=\boldsymbol{\Lambda}_{0}+\boldsymbol{\Theta} t_{i}$, the joint model is then

$$
\mathbf{X}\left(t_{i}\right) \sim \mathcal{P} o\left(\boldsymbol{\omega}_{0} \oplus\left[\left(\boldsymbol{\Lambda}_{0}+\boldsymbol{\Theta} t_{i}\right) \oplus \mathbf{z}\left(t_{i}\right)+\boldsymbol{\lambda}_{b}\right]\right)=\mathcal{P} o\left(\operatorname{diag}\left[\boldsymbol{\omega}_{0}\right] \cdot\left[\left(\boldsymbol{\Lambda}_{0}+\boldsymbol{\Theta} t_{i}\right) \cdot \mathbf{z}\left(t_{i}\right)+\boldsymbol{\lambda}_{b}\right]\right)
$$

or distributing

$$
\begin{aligned}
\mathbf{X}\left(t_{i}\right) & \sim \mathcal{P} o\left(\operatorname{diag}\left[\boldsymbol{\omega}_{0}\right] \cdot \boldsymbol{\Lambda}_{0} \cdot \mathbf{z}\left(t_{i}\right)+\operatorname{diag}\left[\boldsymbol{\omega}_{0}\right] \cdot \mathbf{\Theta} \cdot\left[t_{i} \mathbf{z}\left(t_{i}\right)\right]+\operatorname{diag}\left[\boldsymbol{\omega}_{0}\right] \cdot \boldsymbol{\lambda}_{b}\right) \\
& \sim \mathcal{P} o\left(\mathbf{a}+\mathbf{B} \cdot \mathbf{z}\left(t_{i}\right)+\mathbf{C} \cdot\left[t_{i} \mathbf{z}\left(t_{i}\right)\right]\right) .
\end{aligned}
$$

with intercept vector $\mathbf{a}=\operatorname{diag}\left[\boldsymbol{\omega}_{0}\right] \cdot \boldsymbol{\lambda}_{b}$ and two matrices of coefficients $\mathbf{B}=\operatorname{diag}\left[\boldsymbol{\omega}_{0}\right] \cdot \boldsymbol{\Lambda}_{0}$ and $\mathbf{C}=\operatorname{diag}\left[\boldsymbol{\omega}_{0}\right] \cdot \boldsymbol{\Theta}$. Arrived at this point, it is possible to split the problem in the several components of the response, i.e.

$$
X_{j}\left(t_{i}\right) \sim \mathcal{P}_{o}\left(a_{j}+\mathbf{b}_{j}^{\prime} \cdot \mathbf{z}\left(t_{i}\right)+\mathbf{c}_{j}^{\prime} \cdot\left[t_{i} \mathbf{z}\left(t_{i}\right)\right]\right)
$$

where $\mathbf{a}=\left[a_{1}, a_{2}, \ldots, a_{D}\right]$ and where $\mathbf{b}_{j}^{\prime}$ and $\mathbf{c}_{j}^{\prime}$ are the $j$-th rows of the matrices $\mathbf{B}$ and $\mathbf{C}$ respectively. These models are all estimable with the same GLM framework, with a response of the Poisson family and an identity link. Note that the estimation of this model requires the use of several standards with different compositions.

As we did in the preceding case, in the calibration step we use the data $\mathbf{X}^{s}$ and the known predictors $\left\{t_{i} \in T\right\}$ and $\left\{\mathbf{z}_{k}, k \in \mathcal{K}_{s}\right\}$ to obtain estimates $\tilde{a}_{j}, \tilde{\mathbf{b}}_{j}$ and $\tilde{\mathbf{c}}_{j}$ of the model parameters. In the prediction step, fixing these parameters on their estimated values, the model

$$
X_{j}\left(t_{i}\right) \sim \mathcal{P} o\left(\tilde{a}_{j}+\left[\tilde{\mathbf{b}}_{j}^{\prime}+t_{i} \tilde{\mathbf{c}}_{j}^{\prime}\right] \cdot \mathbf{z}\left(t_{i}\right)\right)
$$

can be fitted with an offset $\tilde{a}_{j}$ and known predictors $\left[\tilde{\mathbf{b}}_{j}^{\prime}+\tilde{\mathbf{c}}_{j}^{\prime} t_{i}\right]$ to obtain estimates of $\mathbf{z}\left(t_{i}\right)$. Note that one can obtain a (different) estimate of the whole vector $\mathbf{z}\left(t_{i}\right)$ for each component of the counts $X_{j}\left(t_{i}\right)$. Alternatively, one can also use the model for the total counts

$$
\mathbf{1}^{\prime} \cdot \mathbf{X}\left(t_{i}\right) \sim \mathcal{P} o\left(\mathbf{1}^{\prime} \cdot \tilde{\mathbf{a}}+\mathbf{1}^{\prime} \cdot\left[\tilde{\mathbf{B}}+t_{i} \tilde{\mathbf{C}}\right] \cdot \mathbf{z}\left(t_{i}\right)\right)
$$

### 4.3 Further models

The preceding models were characterized by an identity link function, a requirement of the additive nature of the blank and the signal. If this condition is removed, then the class of models can be extended to models with logarithmic link function (actually, the canonical choice of Poisson GLMs). This setting would be suitable to consider the multiplicative models mentioned in section 3.4. The same structure as before would then be used, namely a calibration phase in which all parameters would be estimated with a GLM; and a prediction phase in which another GLM with an offset (equal to the multiplicative blank) would be used to estimate the unknown composition of the samples.

All models mentioned before share several main limitations:

- perfectly known standard compositions are required;
- it is not possible to model inhomogeneities of the materials considered (standards or samples);
- no solution exists for other more realistic and flexible models, like a multiplicative interactionperturbation model with additive error;
- if the hypothesis of Poisson distribution is verified to be inaproppriate, the likelihood cannot be computed exactly, and GLM fitting procedures might fail.

All these issues can be tackled with Bayesian estimation techniques. Though they are not much more complex than the methods presented so far, these fall beyond the scope of this contribution and are left for future research. Interested readers can consider the work of van den Boogaart et al. (2013), dealing with a model that can accomodate some of these effects.

## 5 Application

To illustrate the presented concepts, models and solving techniques we use a data set of geochronologically relevant isotopes. These are 35 samples, including 9 analyses of standards, which were analsed for 6 isotopes: Hg202, Pb204, Pb206, Pb207, Pb208, Th232, U235, U238, of which the first two are only used for quality control (Fig. 1). This figure shows 9 samples in 3 blocks of 3 , which composition is very similar: these are actually standards. It is also possible to see that, except for some outliers, the isotopes are ordered in decreasing order of abundance as U238, then Th232 or Pb206, and finally U235, Pb208 and Pb207 in varying orders. Dwell times considered are reported in Table 5

Table 1: Dwell times for the several elements considered

| Hg202 | Pb204 | Pb206 | Pb207 | Pb208 | Th232 | U235 | U238 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1.00 | 1.00 | 0.50 | 2.00 | 1.00 | 0.50 | 0.50 | 0.50 |

A closer look at the last three measurement periods (Fig. 2), including one standard and two samples, shows several remarkable aspects:

- First, Hg 202 and Pb 204 do not show up in the samples or in the standard in a way significantly different from the blank. These elements are used for analytical quality control, and are not relevant for the problem itself.
- Second, several isotopes show clear exponentially decreasing trends while being shot (linear in $\log$ scale), an effect of fractionation of the gas cloud while the laser penetrates deeper in the material.
- Finally, a more relevant aspect for modelling, the variability of these time series is neither additive nor multiplicative, because the background and the signal do not have comparable levels of variability neither in raw nor in $\log$ scales. This is the reason why purely multiplicative models (like those presented in section 3.4) are not realistic.


Figure 1: Time series of counts obtained for each component considered, in raw scale (top) and in log scale (bottom)

Other aspects and relevant concepts are shown as well is Figure 3. Beside the clear fractionation effects occuring while the laser is shot (increasing trend, followed by a shallow exponential decrease) or right after switching it off (pronounced decrease), we see as well the presence of zonations of different composition, thus potentially of different age. The figure also shows that readings have much stronger drifts and variabilities than the relevant isotopic logratios in the measuring windows. A further assessment on the possible structure of the variability of this series is displayed in Figure 4, which makes use of the property that Poisson distributed variates should show dispersion coefficients of $\hat{D}=1$. For the blank windows, overdispersion is clear to see in the heavy ions (U238, Th232). Dispersion coefficients for the measurement period are only reported for the sake of completeness, as they are difficult to estimate due to the presence of the irregularities mentioned before: the values reported in this case are obtained using the residual variance with respect to a fitted exponential trend, which might deliver reasonable estimates for well-behaved samples, but is completely inappropriate for zoned samples. Nevetheless, it shows roughly a similar distribution with a certain bias towards underdispersion.

Given the considerations of this preliminary descriptive analysis, the natural conclusion must be that appropriate models for this dataset should be flexible enough to consider at least the following three effects:

1. natural variability on the composition of the materials (these are not homogeneous), in particular including zonation of samples and random inhomogeneities of standards and samples;
2. downhole fractionation effects of heavy isotopes, which appears as local exponentially de-


Figure 2: Time series of counts obtained for each component considered, in raw scale and in log scale, for the last three samples. In particular, note that the variability of the data has neither a pure additive nor a pure multiplicative structure.


Figure 3: Time series of counts obtained for each component considered, for a well-behaved sample (left) and for a sample showing clear zonation (right). Upper plots show the counts (in log scale) while lower plots show the naive logratios relevant for geochronological calculations. Note as well the dashed vertical lines, showing the windows for blank (between the first two lines) and measurement (between the last two lines).


Figure 4: Dispersion coefficients of the blank and measurement windows for each sample. Vertical dashed lines mark the standards.
creasing trends;
3. Poisson overdispersion (something which is included in state-of-the-art GLM fitting models);
4. additive blank variability.

Unfortunately, none of the models presently implemented within the framework of generalized linear models can deal with all these effects, the main limitation being that one cannot simmultaneously treat as linear the additive background and the exponential trends. Nevertheless, as we have opted for leaving Bayesian methods for future work, the following is an approximate set of results ignoring the first and third shortcomings using a Poisson GLM regression with identity link (Section 4.2.1).

Results (Fig. 5) show several interesting patterns. The proportionality factors for the standards (Fig. 5 lower plot) should be all 1. The standards show a remarkably constant composition, suggesting that no long-range time drift is necessary. The method proposed is quite robust, for instance showing no to minor influences of the outliers of Pb208 in samples 915-005 or Z-025. If we order the elements by decreasing abundance on the estimated composition and on the observed counts, results are the same, which given the several shortcomings we had to take is a good result.

## 6 Discussion

The most obvious implication of these results is the fact that models without additive error cannot be accepted as reasonable descriptions of the physical LA-ICP-MS measurement process, specially when the target value is small because then the contribution of the multiplicative error to the measurement uncertainty becomes irrelevant in comparison with the contribution of the additive error. That was particularly visible on the reduction of the estimated concentration of the minor isotopes of the standards.

This has implications in the treatment of values below the detection limit: for all purposes, the values below the detection limit obtained from mass spectrometry data (at least those obtained with techniques analogous to ICP-MS) should not be treated with an additive logistic normal model, because those dramatically underestimate the uncertainty of the values below and around the detection limit.

A second family of implications relates to the fact that logratios seem to be much more effectively estimated than their numerator and denominator elements separately. Nevertheless, this is not conclusive, and further studies focusing on this aspect are required.

Finally, to obtain a full joint compositional calibration model remains a difficult task because of the need of several standards of different composition, which should be each homogeneous, perfectly known and of the same kind of material than the samples to analyse. Given the practical problems of fulfilling these conditions in real-world applications even for a single standard, it is foreseeable that ICP-MS calibration will remain univariate.

The proposed methodology can nevertheless produce sensible results, even ignoring the multivariate


Figure 5: Results of the analysis of counts for each isotope: (1) original data, (2) results for the major components, (3) results for the minor components, and (4) proportionality factors.
nature of the underlying composition. Though it is premature to extract conclusions of this single study, this is perhaps so because ICP-MS does actually count individual atoms/ions/isotopes, and matter (or mass) is an additive property. Building a composition out of the obtained measurements of the several isotopes is then a choice of the analyst, and not intrinsically demanded by the data. This might not apply to other analytical techniques.

## References

van den Boogaart, K. G.; Tolosana-Delgado, R.; Hron, K.; Templ, M. and Filzmoser, P. (2013) Compositional regression with unobserved components or below detection limit values. In K. Hron, P. Filzmoser and M. Templ (Eds.), Proceedings of CoDaWork'2013: the fifth international Workshop on Compositional Data Analysis, pp. 10-19.
van den Boogaart, K. G. and Tolosana-Delgado, R. (2008) "Compositions": a unified R package to analyze compositional data. Computers and Geosciences, 34 (4), pp. 320-338.

Nelder, J. and Wedderburn, R. (1972) Generalized Linear Models. Journal of the Royal Statistical Society, Series A 135(3), pp. 370-384.

Pawlowsky-Glahn, V. and Egozcue, J.J. (2001) Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research adn Risk Assessment, 15(5), pp. 384-398.

Pawlowsky-Glahn, V., Egozcue, J.J. and Tolosana-Delgado, R. (2015). Modelling and Analysis of Compositional Data, Chichester: Wiley.

Skellam, J. G. (1946) The frequency distribution of the difference between two Poisson variates belonging to different populations. Journal of the Royal Statistical Society, Series A 109(3), pp. 296.

# Variable Selection and Estimation for Regression Models with Compositional Data Predictors 

Huiwen Wang ${ }^{1}$, Lele Huang ${ }^{1 *}$, Liying Shangguan ${ }^{2}$, and Siyang Wang ${ }^{3}$<br>${ }^{1}$ Beihang University, Beijing, CHINA; nanhuabiren@163.com<br>${ }^{2}$ China Galaxy Securities, CO., LTD, Beijing, CHINA<br>${ }^{3}$ Central University of Finance and Economics, Beijing, CHINA


#### Abstract

As the development of computer-related technology, collection and storage of data become more and more convenient, which makes the existence of unrelated variables unavoidable in a specific problem. In this setting, various variable selection methods have been proposed in the literature, but there is little consideration about the structure among the predictors, especially about compositional data. For compositional data, the summation of nonzero components is one, which is a prior sum constraint before modelling and makes the consideration of correlation between components indispensable. In the regression model, when there are multiple compositional data and multivariate scalar data mixed in the predictors for predicting the scalar response, we propose to select the significant predictors and estimate the associated coefficients simultaneously with this mixed type of data in groups for compositional data and individuals for multivariate scalar data, where several group variable shrinkage methods can be implemented in this procedure, such as group Lasso and group SCAD. Besides, considering the robustness of estimator, we choose the quantile loss function to weaken the effect of outliers and reflect the distribution of the reesponse. Simulation and a real data example illustrate the effectiveness of our proposed method.


Key words: group variable selection, compositional data, quantile, norm

## 1 Introduction

Compositional data appear frequently in a wide range of applications, such as topic composition$s$ of documents in machine learning, geochemical compositions of rocks, household patterns in economics and so on. It is the fact that the components of a composition must sum to unity that distinguishes compositional data analysis from other fields or types of data. This constraint makes many traditional multivariate statistical methods be inapplicable to compositional data. The recent years have witnessed the fruitful development of compositional data analysis, since the seminal work of Aitchison (1982). Nowadays, the rapid development of computer-related technology brings big data in both sample size and dimensionality. The research in compositional data analysis concentrates on cases with fixed numbers of components, or the number of compositions is one or much smaller than the sample size.

Lin et al. (2014) proposed variable selection procedure in the framework of regression model with compositional covariates, where the number of components of a composition is diverging with sample size. However, even we obtained the significant components, we have to know the values of all components to attain the percentages. Little formal effort has been made to consider principal tools with diverging number of compositions. Here we concern about regression models with diverging numbers of compositions, but for each compositional predictor, the number of components is fixed. Under this condition, if we want to select significant predictors, they must be selected in groups. For a composition, if at least one components are significant, it can enter the model. Unless all components of a composition are not significant, it is removed from the model.

Besides, in a regression model, it is rare that all the predictors are compositional ones. Usually some scalar predictors appear and play significant roles. For example, in the macroeconomics, to predict the percent of people out of work, the internal structure of the whole economics, such as the percentages of first, second and third production, may be significant. At the same time, the effect of the total amount of gross domestic production (GDP) can not be ignorable. In modeling, both compositional data and common scalar covariates should be included.

A main focus in the literature has been the conditional mean function $E(Y \mid \cdot)$, while, under many circumstances structural change in the conditional quantile function is more important. For example, when studying income inequality, it is more interesting to examine whether (and how) the income differential between different racial groups, conditional on observable characteristics, has changed over time. An increase in inequality may increase the conditional dispersion of the differential, while leaving the mean change little or even unchanged. Thus, the conditional mean ceases to be informative and the conditional quantiles should be considered (Oka and $\mathrm{Qu}, 2011$; Qu and Perron, 2007).

The main purpose of this paper is to identify significant predictors from many predictors, including compositional and common scalar ones under the quantile loss function. During the procedure, group variable selection methods are implemented. Regularization methods (or penalty methods) for simultaneous variable selection and parameter estimation in linear regression and more general contexts have gained intense interest in these years. Fan and Li (2001) proposed a variable selection method via penalized likelihood approaches and summarized many regularization methods.Fan and $\operatorname{Lv}$ (2010) presented a brief account of the recent developments of theory, methods, and implementations for high dimensional variable selection.

Let $Y$ be a real-valued random variable, $\left\{\mathbf{X}_{j}, 1 \leq j \leq p\right\}$ be compositional predictors, and $Z$ is a $q$ - dimensional random vector. It's assumed that the scalar response $Y$ is linearly related to the compositional predictors $X_{j}^{\prime}$ s and multivariate scalar predictors $Z$ through the relationship

$$
\begin{equation*}
Y=\alpha+\sum_{j=1}^{p}\left\langle\beta_{j}, \mathbf{X}_{j}\right\rangle+g(Z)+\epsilon \tag{1}
\end{equation*}
$$

where $Z$ is a random vector and by the unknown function $g(\cdot)$ it affects the response. In fact, $g(\cdot)$ can be usually estimated by spline tools and transformed into linear combination of basis function,
then for simplicity, it is assumed that $g(Z)=Z^{T} \gamma$, then the considering model is

$$
\begin{equation*}
Y=\alpha+\sum_{j=1}^{p}\left\langle\beta_{j}, \mathbf{X}_{j}\right\rangle+Z^{T} \gamma+\epsilon \tag{2}
\end{equation*}
$$

where $\left\{\gamma, \beta_{j}, 1 \leq j \leq p\right\}$ are unknown and need estimating. The $\langle\cdot, \cdot\rangle$ in $\left\langle\beta_{j}, \mathbf{X}_{j}\right\rangle$ is an uniform denotation of inner product operation. $\left\langle\beta_{j}, \mathbf{X}_{j}\right\rangle=\sum_{k=1}^{s_{j}} x_{j k} b_{j k}$, where $\mathbf{X}_{j}=\left[X_{j 1}, \cdots, X_{j s_{j}}\right]$ is a composition with $s_{j}$ components, while $\beta_{j}=\left[b_{j 1}, \cdots, b_{j s_{j}}\right]$ is the corresponding coefficient vector. For simplicity, let $s_{1}=\cdots=s_{j} \equiv s_{0}$. In practice, they can be different. Model (2) can be rewritten as

$$
\begin{equation*}
Y=\alpha+\sum_{j=1}^{p} \sum_{k=1}^{s_{0}} X_{j k} b_{j k}+Z^{T} \gamma+\epsilon \tag{3}
\end{equation*}
$$

## 2 Variable Selection and Estimation

In this section, we will introduce the specific procedure to select significant predictors and estimate the associated coefficients simultaneously, considering the number of covariates may be larger than sample size. The intercept $\alpha$ is omitted here, since it can be deleted by centering the response and covariates.

### 2.1 The linear log-contrast model

The unit-sum constraint renders the $s_{0}$ components of a compositional predictor not vary freely. Therefore, it is necessary to omit certain components to ensure identifiability. However, this operation will bring the difficulties in providing appropriate interpretations for the associated parameters. To resolve this problem, Aitchison and Bacon-shone (1984) proposed the linear logcontrast model by applying the log-ratio transformation to compositional predictors. By the same operation, model (2) can be updated as

$$
\begin{equation*}
Y=\sum_{j=1}^{p} H_{j}^{-} \beta_{j}^{*-}+Z^{T} \gamma+\epsilon \tag{4}
\end{equation*}
$$

where $H_{j}^{-}=\left[\log \left(X_{j k} / X_{j s_{0}}\right), 1 \leq j \leq s_{0}-1\right], \beta_{j}^{*-}=\left[b_{j 1}^{*}, \cdots, b_{j, s_{0}-1}^{*}\right]^{T}$ is the associated $s_{0}-1$ dimensional coefficient vector. Also we can obtain a more conveniently expressed symmetric model by introducing new coefficients $b_{j, s_{0}}^{*}=-\sum_{k=1}^{s_{0}-1} b_{j k}^{*}$ as

$$
\begin{equation*}
Y=\sum_{j=1}^{p} H_{j} \beta_{j}^{*}+Z^{T} \gamma+\epsilon, \quad \sum_{k=1}^{s_{0}} b_{j k}^{*}=0,1 \leq j \leq p \tag{5}
\end{equation*}
$$

where $H_{j}=\left[\log \left(X_{j 1}\right), \cdots, \log \left(X_{j s_{0}}\right)\right], \beta_{j}^{*}=\left[b_{j 1}^{*}, \cdots, b_{j, s_{0}}^{*}\right]^{T}$.
Denote $\tilde{H}=\left[H_{1}, \cdots, H_{p}, Z\right], B^{*}=\left[\beta_{1}^{* T}, \cdots, \beta_{p}^{* T}, \gamma^{T}\right]^{T}$, then our model can be expressed as

$$
\begin{equation*}
Y=\tilde{H}^{T} B^{*}+\epsilon, \quad \text { subject to } \tag{6}
\end{equation*}
$$

$$
R B^{*}=\left(\begin{array}{cccccccccccc}
1 & \cdots & 1 & 0 & \cdots & \cdots & & \cdots & 0 & 0 & \cdots & 0 \\
0 & \cdots & 0 & 1 & \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & & & & 0 & \cdots & 0 \\
0 & \cdots & 0 & \cdots & \cdots & \cdots & 1 & \cdots & 1 & 0 & \cdots & 0
\end{array}\right) B^{*}=W=\left(\begin{array}{l}
0 \\
0 \\
0 \\
0
\end{array}\right)
$$

It is clear that this is a regression model with linear constraints on the coefficients. By Lagrangian multiplier method, it can be obtained that

$$
\begin{equation*}
\hat{B}^{*}=B_{L S}-\left(\tilde{H}^{T} \tilde{H}\right)^{-1} R^{T}\left[R\left(\tilde{H}^{T} \tilde{H}\right)^{-1} R^{T}\right]^{-1}\left[R B_{L S}-W\right], \tag{7}
\end{equation*}
$$

where $B_{L S}$ is the ordinary least square estimator without constraints. However, the estimator (7) is based on the assumption that all predictors are significant for the response, which is usually not true. Then we need distinguish the significant ones from redundant variables.

## 3 Group variable selection

Assume we have obtained the observations $\left.\left\{\left(y_{i}, z_{i}, x_{i, j 1}, \cdots, x_{i, j s_{0}}\right), 1 \leq j \leq p, 1 \leq i \leq n\right)\right\}$. Based on model (5), we consider the constrained convex optimization problem to attain the estimator

$$
\begin{equation*}
\left(\hat{\beta}^{*-}, \hat{\gamma}\right)=\arg \min \sum_{i=1}^{n} \rho_{\tau}\left(y_{i}-\sum_{j=1}^{p} H_{i j}^{-} \beta_{j}^{*-}-z_{i}^{T} \gamma\right)+\sum_{j=1}^{p} \lambda_{j}\left\|\beta_{j}^{*-}\right\|_{\infty}+\sum_{j=1}^{q} \mu_{j}\left|\gamma_{j}\right| \tag{8}
\end{equation*}
$$

where $\rho_{\tau}(t)=t(\tau-I(t<0)), H_{i j}^{-}=\left[\log \left(X_{i, j k} / X_{i, j s_{0}}\right), 1 \leq k \leq s_{0}-1\right], \beta_{j}^{*-}=\left[b_{j 1}^{*}, \cdots, b_{j, s_{0}-1}^{*}\right]^{T}$, $\|\cdot\|_{\infty}$ means the sup-norm and $\left\|\beta_{j}^{*-}\right\|_{\infty}=\max \left\{\left|b_{j 1}^{*}\right|, \cdots,\left|b_{j, s_{0}-1}^{*}\right|\right\}$.
From the above formulation, we only need to solve a linear program, which greatly simplifies the computation. Here we take the advantage of the specific choice of loss function and penalty function. As we have mention early, other losses and penalty functions can also be chosen in principle. However, complicate penalties and loss functions usually involves intensive computation.

In the penalty term we choose the sup-norm, and $\left\|b_{j}\right\|_{\infty}<C$ is equivalent to the every compoent $b_{j k}<C, k=1, \cdots, s_{0}$. Similar to Bang and Jhun (2012), the optimization problem (8) can be reformulated as a linear programming problem by introducing $2 n$ slack variables and then lpsolve package provided in R can be used to implement this linear programming problem.

Denote the the slack variable by $\left\{\left(u_{i}, v_{i}\right), i=1, \cdots, n\right\}$ and they represent the positive and negative parts of residuals for the $i$ th observation. They satisfy the equality constraint

$$
y_{i}-\sum_{j=1}^{p} H_{i j}^{-} \beta_{j}^{*-}-z_{i}^{T} \gamma=u_{i}-v_{i} .
$$

Also $b_{j k}$ can be rewritten as $b_{j k}=b_{j k}^{+}-b_{j k}^{-}$, then minimizing the objective function in (8) can be obtained by minimizing

$$
\frac{1}{n} \sum_{i=1}^{n}\left(\tau u_{i}+(1-\tau) v_{i}\right)+\sum_{j=1}^{p} \lambda_{j} M_{1 j}+\sum_{l=1}^{q} \mu_{l} M_{2 l}
$$

subject to

$$
\begin{aligned}
& \sum_{j=1}^{p} H_{i j}^{-}\left(b_{j k}^{+}-b_{j k}^{-}\right)+\sum_{l=1}^{q} z_{i l}\left(\gamma_{l}^{+}-\gamma_{l}^{-}\right)+u_{i}-v_{i}=y_{i} \\
& M_{1 j} \geq b_{j k}^{+}+b_{j k}^{-}, \quad M_{2 l} \geq \gamma_{l m}^{+}+\gamma_{l m}^{-} \\
& b_{j k}^{+} \geq 0, \quad b_{j k}^{-} \geq 0, \quad \gamma_{l m}^{+} \geq 0, \gamma_{l m}^{-} \geq 0, \quad u_{i} \geq 0, \quad v_{i} \geq 0
\end{aligned}
$$

for $i=1, \cdots, n, j=1, \cdots, p, l=1, \cdots, q$.
From the above formulation, we only need to solve a linear program, which greatly simplifies the computation. Here we take the advantage of the specific choice of loss function and penalty function.

### 3.1 Tuning Parameter

In (8) we can also utilize other norm. In Yuan and Lin (2006), they chose to penalize the $L_{2}$ norm of the coefficients within each group. Note that both $\left\|\beta_{j}^{*-}\right\|_{\infty}$ and $\left\|\beta_{j}^{*-}\right\|_{2}$ are singular only when
the whole vector $\beta_{j}^{*-} \equiv 0$, then the $L_{2}$ norm penalty and $L_{\infty}$ norm penalty select variables in groups, or "all-in-all-out" fashion.
Let $\lambda_{j}=\lambda\left\|\tilde{b}_{j}\right\|_{\infty}^{-1}, \mu_{j}=\lambda\left\|\tilde{\gamma}_{j}\right\|^{-1}$ (where $\tilde{b}_{j}, \tilde{\gamma}_{j}$ are the unpenalized quantile estimator), then only $\lambda$ needs selecting. As to the selection of regularization parameter $\lambda$ we use the generalized information criterion proposed by Fan and Tang (2013). we define the generalized information criterion in our problem as

$$
\begin{equation*}
G I C(\lambda)=\log \hat{\sigma}_{\lambda}^{2}+\left(s_{\lambda}+q-1\right) \frac{\log \log n}{n} \log ((p+q) \vee n) \tag{9}
\end{equation*}
$$

where $\hat{\sigma}^{2}=\sum_{i=1}^{n}\left(y_{i}-\hat{y}_{i}\right)^{2} / n, s_{\lambda}$ is the number of groups with nonzero coefficients and $(p+q) \vee n=$ $\max (p+q, n)$. Alternatively, cross validation methods can be implemented, such as the $k-$ fold cross validation, which can trade off between model selection consistency and prediction accuracy.

## 4 Large Sample Properties

## 5 Numerical Results

### 5.1 Simulation

We conducted simulation studies to compare the numerical performance of the proposed estimator.
Similar to Lin et al. (2014), we generated the covariate data $X_{1}, X_{2}, X_{3}$ in the following way, respectively. We first generated an $n \times s_{0}$ data matrix $W=\left(\omega_{i j}\right)$ from a multivariate normal distribution $N_{s_{0}}(0, \Sigma)$, then by the transformation $x_{i j}=\exp \left(\omega_{i j}\right) / \sum_{k=1}^{s_{0}} \exp \left(\omega_{i k}\right)$ we can obtain the covariate matrix $X=\left(x_{i j}\right)$. To describe different levels of correlations among components, let $\Sigma_{s t}=0.3^{|s-t|}, 1 \leq s, t \leq s_{0}$. Set $s_{0}=5, \beta_{1}=[1,0,-1,0.5,0], \beta_{2}=\beta_{3}=0$. We trichotomize $Z_{1}$ as 0,1 or 2 by the quantiles $1 / 3,2 / 3$ of standard normal distribution. $Z_{2}$ is valued in $\{0,1\}$ with the same probability. $Z_{3}$ is also from the standard normal distribution. Then true response variables are generated by

$$
\begin{equation*}
Y_{i}=\left\langle X_{1 i}, \beta_{1}\right\rangle+0.05 Z_{1 i}+0.02 I\left(Z_{2 i}=0\right)-0.03 Z_{3 i}+\sigma \epsilon_{i} \tag{10}
\end{equation*}
$$

The following three different distributions are considered for $\epsilon_{i}$ : the standard t-distribution with 3 degree freedom $(T(3))$, the standard t-distribution with 5 degree freedom $(T(5))$, the standard double exponential (dexp) and Cauchy distribution. Two different values are tested for $\sigma 0.1$ and 0.3 , representing weak and strong signal-to-noise ratios. For comparison, results by group lasso method with square loss function is listed. "Q" and "G" mean the proposed quantile method in this paper and group lasso method with square loss function, respectively. Average number of correctly identified nonvanishing coefficients (TP) and average number of incorrectly identified nonvanishing coefficients (FP) are listed.

Table 1: Simulation results with $T(3)$ error with $\sigma=0.1$

| Q.TP |  |  |  |  |  |
| ---: | ---: | ---: | ---: | ---: | ---: |
| $\tau=0.25$ | Mean | 3.8920 | 3.9230 | 1.3700 | 1.3600 |
|  | Sd | 0.3451 | 0.2727 | 0.9391 | 0.9972 |
| $\tau=0.50$ | Mean | 3.8600 | 3.9900 | 0.8700 | 1.8000 |
|  | Sd | 0.3835 | 0.1000 | 0.8246 | 0.9640 |
| $\tau=0.75$ | Mean | 3.9100 | 3.9300 | 2.1100 | 1.8700 |
|  | Sd | 0.2727 | 0.2564 | 1.0411 | 0.9812 |

According to Table 1-6, we can know that quantile method and group lasso method can select the nonzero scalar and compositional predictors correctly, but they tend to select predictor with zero coefficients into the model. When $\tau=0.50$, the quantile selection result is better than group lasso.

Table 2: Simulation results with $T(3)$ error with $\sigma=0.3$

|  |  | Q.TP | G.TP | Q.FP | G.FP |
| :--- | ---: | ---: | ---: | ---: | ---: |
| $\tau=0.25$ | Mean | 3.8900 | 3.7400 | 1.8200 | 1.4100 |
|  | Sd | 0.3145 | 0.5140 | 1.0385 | 0.9650 |
| $\tau=0.50$ | Mean | 3.8300 | 3.8700 | 1.0000 | 1.4500 |
|  | Sd | 0.3775 | 0.3487 | 0.7785 | 0.8919 |
| $\tau=0.75$ | Mean | 3.8400 | 3.7200 | 1.7900 | 1.6600 |
|  | Sd | 0.4197 | 0.4794 | 1.0306 | 0.9767 |

Table 3: Simulation results with $T(5)$ error with $\sigma=0.1$

|  |  | Q.TP | G.TP | Q.FP | G.FP |
| ---: | ---: | ---: | ---: | ---: | ---: |
| $\tau=0.25$ | Mean | 3.9760 | 3.9430 | 1.9200 | 1.7100 |
|  | Sd | 0.2227 | 0.2387 | 0.9917 | 1.0080 |
| $\tau=0.50$ | Mean | 3.8900 | 3.9600 | 1.1100 | 1.7200 |
|  | Sd | 0.3145 | 0.2429 | 0.7233 | 0.8885 |
| $\tau=0.75$ | Mean | 3.9300 | 3.9800 | 1.8700 | 1.6500 |
|  | Sd | 0.2564 | 0.1407 | 0.8487 | 0.8805 |

Table 4: Simulation results with $T(5)$ error with $\sigma=0.3$

|  |  | Q.TP | G.TP | Q.FP | G.FP |
| ---: | ---: | ---: | ---: | ---: | ---: |
| $\tau=0.25$ | Mean | 3.8850 | 3.8370 | 1.8600 | 1.4000 |
|  | Sd | 0.3562 | 0.4034 | 0.9103 | 0.8762 |
| $\tau=0.50$ | Mean | 3.8740 | 3.8020 | 1.2400 | 1.4900 |
|  | Sd | 0.3380 | 0.4020 | 0.8055 | 0.9999 |
| $\tau=0.75$ | Mean | 3.8760 | 3.8510 | 1.9300 | 1.4500 |
|  | Sd | 0.3667 | 0.3860 | 0.9771 | 1.0577 |

Table 5: Simulation results with double exponential error with $\sigma=0.1$

|  |  | Q.TP | G.TP | Q.FP | G.FP |
| ---: | ---: | ---: | ---: | ---: | ---: |
| $\tau=0.25$ | Mean | 3.8900 | 3.9200 | 1.8800 | 1.4600 |
|  | Sd | 0.3451 | 0.2727 | 0.9773 | 0.9579 |
| $\tau=0.50$ | Mean | 3.9000 | 3.9000 | 1.0300 | 1.6900 |
|  | Sd | 0.3015 | 0.3015 | 0.7029 | 0.9713 |
| $\tau=0.75$ | Mean | 3.8900 | 3.8700 | 1.9000 | 1.6700 |
|  | Sd | 0.3015 | 0.3380 | 0.8469 | 0.9434 |

Table 6: Simulation results with double exponential error with $\sigma=0.3$

|  |  | Q.TP | G.TP | Q.FP | G.FP |
| :--- | ---: | ---: | ---: | ---: | ---: |
| $\tau=0.25$ | Mean | 3.8100 | 3.6200 | 2.1200 | 1.3600 |
|  | Sd | 0.4264 | 0.5646 | 0.8200 | 0.8821 |
| $\tau=0.50$ | Mean | 3.7300 | 3.6100 | 1.3400 | 1.3800 |
|  | Sd | 0.4894 | 0.5318 | 0.7551 | 0.9077 |
| $\tau=0.75$ | Mean | 3.8800 | 3.5700 | 2.0000 | 1.4300 |
|  | Sd | 0.3562 | 0.6416 | 0.9211 | 0.8558 |

Table 7: Simulation results with Cauchy error with $\sigma=0.1$

| Table 7: Simulation results with Cauchy error with $\sigma=0.1$ |  |  |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: |
| $\tau=0.25$ |  | Q.TP | G.TP | Q.FP | G.FP |
|  | Sd | 3.7900 | 2.3800 | 1.7300 | 1.0400 |
| $\tau=0.50$ | Mean | 3.83300 | 1.3241 | 0.8147 | 0.9092 |
|  | Sd | 0.4025 | 1.4017 | 1.0300 | 0.9000 |
| $\tau=0.75$ | Mean | 3.7800 | 2.1600 | 1.8000 | 0.8469 |
|  | Sd | 0.4522 | 1.3228 | 0.8876 | 0.9260 |


| Table 8: Simulation results with Cauchy error with $\sigma=0.3$ |  |  |  |  |  |
| :---: | ---: | ---: | ---: | ---: | ---: |
|  |  | Q.TP | G.TP | Q.FP | G.FP |
| $\tau=0.25$ | Mean | 3.5600 | 1.1900 | 2.3500 | 0.8200 |
|  | Sd | 0.6541 | 1.0318 | 0.8333 | 0.9032 |
| $\tau=0.50$ | Mean | 3.6100 | 1.1800 | 1.4000 | 0.7400 |
|  | Sd | 0.5464 | 1.1044 | 0.7521 | 0.8947 |
| $\tau=0.75$ | Mean | 3.6200 | 1.2400 | 2.1300 | 0.7900 |
|  | Sd | 0.5505 | 1.0929 | 0.8487 | 0.8680 |

Especially when the error term are with heavier tails, such as Cauchy distribution, the group lasso can not identify the nonzero predictors effectively, see Table 7,8. The signal-to-noise ration also affects the variable selection result clearly. With higher noise, it is more difficult to identify the true model.

### 5.2 Real Data

## 6 Conclusion

For compositional data, the prior sum constraint makes the modeling more complicated. of correlation between components indispensable. In the paper we considered the regression model with are multiple compositional data and multivariate scalar data mixed in the predictors, and proposed to select the significant predictors and estimate the associated coefficients simultaneously in groups for compositional data and individuals for multivariate scalar data, where several group variable shrinkage methods can be implemented in this procedure. Besides, we concern more about the quantile function, not just the conditional mean. Simulation and a real data example illustrate the effectiveness of our proposed method.

## REFERENCES

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). Journal of the Royal Statistical Society. Series B (Methodological) 44(2), 139-177.

Aitchison, J. and J. Bacon-shone (1984). Log contrast models for experiments with mixtures. Biometrika 71 (2), 323-330.

Bang, S. and M. Jhun (2012). Simultaneous estimation and factor selection in quantile regression via adaptive sup-norm regularization. Computational Statistics \& Data Analysis 56(4), 813-826.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association 96(456), 1348-1360.

Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space. Statistica Sinica 20(1), 101-148.

Fan, Y. and C. Y. Tang (2013). Tuning parameter selection in high dimensional penalized likelihood. Journal of the Royal Statistical Society Series B 75(3), 531-552.

Lin, W., P. Shi, R. Feng, and H. Li (2014). Variable selection in regression with compositional covariates. Biometrika 101(4), 785-797.

Oka, T. and Z. Qu (2011). Estimating structural changes in regression quantiles. Journal of Econometrics 162(2), 248-267.

Qu, Z. and P. Perron (2007). Estimating and testing structural changes in multivariate regressions. Econometrica 75(2), 459-502.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 68(1), 49-67.

# Multi element geochemical modelling for mine planning: Case study from an epithermal gold deposit. 

N. Caciagli Warman

Kinross Gold Corporation, Toronto, Ontario, Canada; Natalie.warman@Kinross.com


#### Abstract

Mineralization and alteration processes will result in zones with distinct geochemical characteristics within an orebody. To visualize the mine scale variability that arises as a result of these processes, geochemical domains are defined using a $k$-means clustering algorithm to analyze multi-element data. The fact that the chemical values can be grouped in a defined 3D location clearly suggests that the clusters have meaning in terms of geological process. Principal component analysis (PCA) of these clusters can further improve the understanding of this variability. These clusters form the basis of the geochemical domains which have direct implications for characterization and proportional sampling of geometallurgical and waste rock domains.

In the case study presented, pre-mining geochemical characterizations were undertaken at an epithermal gold deposit to support metallurgical sampling and mine planning. $k$-means cluster analysis and principal component analysis of the geochemical clusters was used to support the metallurgical sampling program by identifying domains for variability testing. The geochemical clusters identified were used to define the oxide, sulphide and transition zones, a critical factor for mineral processing and recoveries and a key variable in the economics of the project.

The R software environment for statistical computing was used for exploratory data analysis (e.g. PCA; zCompositions, robCompositions) and $k$-means analysis (fpc).


## 1 Introduction

A new gold or base metal mine can take $10-15$ years to permit and construct and involves an investment of $>\$ 1 B$ USD. Inadequate characterization of the orebody can lead to unexpected issues with geotechnical stability, metallurgical recoveries, environmental impacts, or other problems. Kinross has determined that a thorough understanding of its orebodies is critical to the success of these major, long term investments and that geochemistry is a key component of this understanding.

Mineralizing and alteration processes will result in zones with distinct geochemical characteristics. The fact that the chemical values can be grouped in a defined 3D location clearly suggests that the clusters have meaning in terms of geological process. This forms the basis of the geochemical domains, which have direct implications for sampling locations for variability testing of geometallurgical recoveries.

The geochemical domains will reflect the mineralogy of the domains with a granularity not always visible by geological mapping and logging alone. For example, alteration events associated with gold mineralization can overprint the host lithology to a degree that the original rock type is unable to be properly identified, or metasomatism may result in chemical changes to mineralogy not fully evident through visual inspection alone. The use of geochemical domains supports a more concrete link between mineralogy and metallurgical variables such as cyanide consumption, metal recoveries, concentration of deleterious elements, as well as comminution properties (hardness, grindability).

## 2 Methodology

The ultimate goal of the geochemical modelling process is to produce geochemical domains that are relevant for metallurgical study on a scale compatible with the proposed mining method. As such there will be a trade-off between granularity and complexity of the model. Geochemical domains need to be viewed within the context of lithology and alteration so that metallurgical variables can be linked to mineralogy.

### 2.1 Data pre-processing

The dataset used for the geochemical modelling must consist of spatially co-located data. For example; samples for gold analysis and multi elements analysis should be collected over the same intervals. Any hyperspectral logging (qualitative mineralogy) or QEMSCAN (quantitative mineralogy) should also be spatially co-located with the geochemical and metal assay data. The sampling strategy needs to be assessed for possible sampling bias. This can be accomplished by visualization of the data in 3D space, examination of sample lengths and counts by lithology, alteration, structure through histograms and box and whisker diagrams. Exploratory data analysis of the multi element geochemistry (i.e. summary stats, detection limits) is key to understanding the nature and distribution of the dataset (Grunsky, 2010).


Figure 1. Quantile-Quantile plot of imputed values and renorted values for Hg

### 2.1.1 Data imputation

The modified "Expectation-Maximisation" (EM) algorithm from the zCompositions package (Palarea-Albaladejo and MartinFernandez, 2014) within the R statistical computing program was used to impute values below the lower limits of detection (LOD) as well as any missing values in the dataset. This procedure fills in the lower tail of the distribution of each element while still preserving the covariance structure of the data. The zCompositions modified EM-imputation algorithm imputes values based on a reference dataset (i.e. samples with no missing observations) and on the threshold values (i.e. the limit of detection for a given element). Following the application of the imputation
routine every imputed value was visually inspected on a quantile - quantile plot (Figure 1) for a given element to assess the appropriateness of the replacement.

### 2.1.1 Data transformation

The working dataset was transformed using log ratio normalization (Aitchison, 1986). Compositional data is by definition constrained to a constant sum (e.g. reported as percent or per mil) and as such the individual variables will not vary independently. This induced correlation introduces a bias into the covariance structure of the data and may obscure the true relationships between the variables. Log ratio normalization takes into account the constant sum constraint of compositional data and centres the data in a way that removes this bias. For robust PCA (see below) the working datasets are normalized using an isometric log-ratio transform (ilr; Egozcue et al., 2003). K-means clustering is performed on a centred log ratio transformed (clr) dataset.

## 2.2 rPCA

In a mining environment understanding the behaviour of outliers is of significant interest. Gold mineralization tends to be inherently non-uniform and as a result many samples within the ore zone may have "anomalous" values compared to the waste rock. These samples need to be understood with respect to their geological context (lithology, alteration and structure) for vectoring and targeting purposes, as well as for understanding the nature of the gold distribution within the ore body. Unless sampling issues or analytical errors are confirmed the dataset is examined inclusive of outliers.

Principal component analysis (PCA) is the simplest of the true eigenvector-based multivariate analyses. Its purpose can often be thought of as revealing the internal structure of the data in a way which best explains the data variance. It is a technique used to change a set of original variables into a number of basic dimensions. The algorithm used here for the calculation of the principal components was the "Robust PCA" from the robCompositions package (Filmozer et al., 2011) within the R statistical computing program. Robust PCA (rPCA) is less affected by outliers in the dataset and provides more reliable calculation of the covariance matrix for the analysis. The rPCA algorithm from the robCompositions package utilizes a minimum covariance estimator (MCD) by examining a subset of $h$ observations with the smallest determinant of their sample covariance determinant. In order to maximize the robustness of the MCD location the $h$ used here was $1 / 2$ of the total sample size because the number of outliers tends to be high.

## 2.3 k-means cluster analysis

K-means clustering is a method of cluster analysis which aims to partition $n$ observations into $k$ clusters, in which each observation belongs to the cluster with the nearest mean. $K$-means clustering uses the minimum Euclidian distance (difference between values) as the main criterion to discriminate between different groups. These clusters can be back-coded into the database and viewed in 3D space for interpretation within the context of lithology, alteration and metallurgical responses.


Figure 2. An example of a sum of squared errors (SSE) plot versus number of clusters.

The number of clusters to partition a dataset can be assessed using a plot of the sum of squared errors (SSE) versus number of clusters (Figure 2). The best number of clusters to describe the structure in the dataset can be determined by locating the break in slope on a plot of the SSE versus number of clusters. For geochemical modelling of an orebody a combination of the SSE plot, visualization of the clusters in either principal component space or classical discriminant coordinates and a visual (3D) inspection of the clusters in a geological modelling software was used
to determine the "best" number of clusters. For example, from figure 2, four clusters are suggested; however five clusters showed distinct separation in 3D space (see Figure 9).

### 2.3.1 Statistical validation of clusters

The robustness of the clusters was determined using a bootstrapping algorithm from the fcp package (Hennig, 2007) within the R statistical computing program which resamples the dataset 100 times and returns a measure of the stability of the cluster, the Jaccard similarity value, which is assigned a value between 0 and 1 . Generally, a valid, stable cluster should yield a mean Jaccard similarity value of 0.75 or more (Hennig, 2007).

### 2.3.2 Geochemical and spatial validation of clusters



Figure 3. Plot of PC1 vs PC2 coordinates for samples colour coded by cluster number

The geochemical variability of the clusters can be examined within either principal component space or with a discriminant analysis. The sample dataset was tagged with the cluster number and plotted with their rPCA coordinates (see Figure 3). This type of plot can also provide as assessment as to the suitability of the number of clusters selected for $k$-means cluster analysis. For example, samples from cluster 7 (red) and cluster 5 (blue) completely overlap in PC1-PC2 space, suggesting that there is very little variability between these clusters. These samples were then examined spatially to determine if they have a defined location in space (Figure 4) and can be explained within the context of known lithology, alteration or structure or whether they just describe the inherent geochemical heterogeneity within that domain.


Figure 4 Drill hole intervals back-coded with cluster number viewed in 3D space

### 2.4 Wireframes and characterization of domains

The first pass wireframe construction process uses Leapfrog geological modelling software to create surfaces based on the criteria outlined above to encapsulate the back coded clusters. The surfaces are created independently of the lithological or alteration domains, however the geochemical domains will correspond to a combination of the lithological, alteration and weathering domains determined from visual logging.

## 3 Case Study 1

Deposit A is interpreted as having been emplaced into a phreatomagmatic diatreme-dome complex with local diatreme fill. Mineralization is associated with high sulphidation vuggy silica with advanced argillic quartz-alunite or quartz-pyrophyllite alteration which grades outward into quartzclays. The alteration assemblage appears to have a high percentage of silica and lesser amounts of clay. In oxidized portions, the gold and silver are residual, but at depth gold and silver are associated with multiple sulphides (mostly enargite and pyrite, with lesser covellite and sphalerite) and sulphosalts.


Figure 5. Elemental distribution of raw data - all expressed as ppm, (n) indicates samples >LOD

The depth of oxidation varies from less than 100 metres to over 300 metres and is typically around 120-150 metres. There is a narrow and discontinuous "mixed" horizon at the oxide-sulphide interface (transition zone), less than 10 metres thick, but it invariably includes secondary chalcocite.

Within this transition zone metal recoveries can be significantly less than in the oxide zone; in this case $75 \%$ recovered Au compared to $82 \%$. A robust definition and identification of this domain is critical to determining the economics of the project. However determination of this transition zone by visual logging is difficult and subjective.

A detailed geochemical model was created within the mineralised vuggy silica zone of Deposit A to determine a consistent and unbiased definition of the oxide, transition and sulphide zones to accurately assess the amount of recoverable metals and better understand the economic potential of the deposit.

### 3.1 Data and Data Pre-Processing

The working dataset consisted of information collected on two very different resolutions. The logged lithology, logged alteration and metal assay data were collected with a 2 metre resolution. In contrast, the multielement data (collected by previous operators) consisted of 10 metre composites created by taking a 20 g sample from every 5 assay sample pulps (originally collected on 2 metre intervals) down hole, homogenizing
those 5 samples and sending an aliquot of the composite for aqua regia digest and multi-element analysis by ICP. In order to properly merge the two datasets the data in the 2 metre $\mathrm{Au}, \mathrm{Cu}, \mathrm{Ag}, \mathrm{As}$, $\mathrm{Pb}, \mathrm{Zn}$ and Hg assay tables needed to be composited over 10 metres. Multi element data sampled over 2 metre intervals, consistent with the Au sampling protocol would be preferable.

The statistical distribution of each element was inspected individually. Elements were removed from the dataset if a significant portion of samples measured below detection limits ( $>50 \%$ ) or if the data was severely quantized due to the analytical resolution. This resulted in the removal of $\mathrm{Sc}, \mathrm{Cd}, \mathrm{Tl}, \mathrm{Ti}$, Th, B, Be, W, La, Ga and U (Figure 5).

## 3.2 rPCA

The Deposit A rPCA was constructed using the multi-element data to examine relationships found in Au-bearing lithology and alteration types. The numerical variables analysed were the assayed elements: Au, Ag, As, Al, Ba, Bi, Ca, Co, Cr, Cu, Fe, Hg, K, Mg, Mn, Mo, Na, Ni, P, Pb, Sb, Sr, Stotal, Sulphide, Sulphate, V and Zn.


Figure 6 rPCA variables map for Deposit A multi element data

Partial digestion, using aqua regia, was used to extract elements for analysis. This method dissolves minerals selectively and therefore the bulk rock chemistry is not reflected in the major elements determined by this method. The recovery of each element will depend on the mineralogy. The aqua-regia method only targets the trioctahedral silicates (biotite, chlorite), clays, sulfosalts and some oxides. It will not recover elements hosted by other silicates and refractory oxides. Consequently, extra care must be given to interpretation of the major element data obtained by partial rock dissolution.

The observed chemical relationships of selected elements show patterns that are assumed to be a function of the type of mineralization and accompanying alteration. In the rPCA variable map in Figure 6 a number of elements are correlating with Au $\left(\mathrm{Cu}, \mathrm{Ag}, \mathrm{Hg}, \mathrm{Fe}, \mathrm{SO}_{4}\right.$ and S ; the vectors are in close proximity to each other). The presence of enargite $\left(\mathrm{Cu}_{3} \mathrm{AsS}_{4}\right)$ is suggested by the Cu -As association. The strong correlation of Al-K implies an illite ( $\mathrm{K}-\mathrm{Al}$ clay) association. The vector describing Zn is the longest vector in the PCA suggesting that the occurrence of sphalerite ( ZnS ) is highly variable; however as demonstrated in the $k$-mean analysis, discussed below, the Zn variability can be spatially defined.

## 3.2 k-means cluster analysis

All the elements used in the rPCA are also used in the clustering exercise. For Deposit A, the preliminary cluster analysis resulted in domains that broadly correspond to the alteration zones (see Figure 4). To provide further resolution of the variability within only the mineralized ore zone, a $k$-means cluster analysis was carried out on the geochemical data from intervals in the vuggy silica oxide, vuggy silica sulphide and alunite alteration zone.

For Deposit A, a combination of the SSE plot, visualization of the clusters in discriminant coordinate space (Figure 8) and 3D space (Figure 9) was used to determine the appropriate number of clusters.

From figure 7, four clusters are suggested; however five clusters showed distinct separation in both geochemical space (Figure 8) and 3D space (Figure 9).


Figure 7 Sum of squared errors (SSE) vs number of clusters


Figure 8 Visualization of the clusters in geochemical space


Figure 9 Deposit A drill hole intervals back-coded with cluster number viewed in 3D space

The robustness of the clusters was determined using a bootstrapping algorithm which resamples the dataset 100 times and returns a measure of the stability of each cluster, the Jaccard similarity value, which is assigned a value between 0 and 1 . Generally, a valid, stable cluster should yield a mean Jaccard similarity value of 0.75 or more. The mean Jaccard values for Clusters 2, 4 and 5 range from 0.82 to 0.84 , while the Jaccard values for Cluster 1 and Cluster 3 are 0.66 (Figure 8 and 9).

These clusters can be briefly characterized as:
Cluster 1: A low copper (mean content of $0.5 \%$ ) high zinc domain within the sulphide zone corresponding to sphalerite rich zones. The gold within this domain is highly refractory and not recoverable by current processing methods.

Cluster 2: A "mixed" horizon at the oxide-sulphide interface with elevated soluble copper content corresponding to the transition domain.

Cluster 3: A high copper (mean content of 2\%) and arsenic domain within the sulphide zone corresponding to enargite rich zones. Due to the high arsenic content and the refractory nature of the gold the metals in this domain are not economically recoverable by current processing methods.

Cluster 4: An oxide domain (mean S content of $0.2 \%$ ) within the vuggy silica zone corresponding to the oxide zone. This domain has the highest gold recoveries.

Cluster 5: A potassium and aluminium rich domain that slightly overprints the vuggy silica alteration and transitions into the quartz-illite alteration halo on the margin of the deposit. There is recoverable metal in this domain, but the differing mineralogy will result in differing behaviour in the mill. This domain also contains the highest mean Pb content of all the domains.

Clusters 2 (transition domain) and 4 (oxide domain) occur predominantly within the portion of the deposit logged as oxide. Visual logging frequently misidentifies transition and oxide materials, particularly near the oxide-sulphide interface due to difficulty in visually identifying these zone and subjective nature of the geological logging (Figure 10). Because the identification of the transition zone is relevant for metallurgical processing and key to determining metal recoveries, this geochemical determination provides a unbiased and consistent way to identify this zone.


Figure 10 Drill core photos of intervals classified as oxide domain (cluster 4) and transition domain (cluster 2). Within the oxide zone vugs are either empty or filled with quartz (A). Within the transition domain vugs are filled with quartz + fine grained disseminated sulphides (photo is cm scale).

The transition domain at the oxide-sulphide interface includes the secondary copper minerals, chalcocite. Chalcocite provides high concentrations of soluble copper will have a negative effect on the total cyanide consumption and possibly reduced gold recoveries.

Figure 11 shows how the spatial separation of clusters 2 and 4 relate to the redox limits as established by visual logging. Chemical cluster 4, interpreted as the oxide domain, is hosted by the vuggy silica alteration well within the visually logged oxide zone while the intervals assigned to cluster 2 , interpreted as transition domain, correspond to the vuggy silica alteration at the interface between the oxide and sulphide zones.


Figure 11 A vertical cross section through the ore zone showing the lithological and spatial relationship of the geochemically determined oxide zone (blue intervals) and transition zone (red intervals). Also shown is the logged vuggy silica domain (light grey shape) and the redox boundary identified by visual logging (darker grey).

Clusters 1 and 3 occur within the visually logged sulphide zone and differ from each other in copper content. Cluster 3 has mean copper content of $2 \%$ and cluster 1 has a mean copper content of $0.5 \%$. Cluster 3 most likely represents copper rich feeder zones or copper rich veins within the low copper domain (cluster 1; Figure 12). However, the overlap between cluster 1 and 3, spatially and geochemically likely reflects the inability of the 10 metre composited data to precisely discriminate at this scale. Depending on the ultimate metallurgical processing flowsheet, resolution of these high copper zones may or may not be necessary.
 sulphide veins. (photo is cm scale)

The high copper domain within the sulphide zone of vuggy silica alteration (cluster 3) is centered on the main feeder zones of the vuggy silica alteration. The low copper domain (cluster 1) is more predominant on the peripheries of the vuggy silica alteration zone to the north and south (Figure 13).


Figure 13 Plan view of the logged vuggy silica alteration (grey shape) within the sulphide zone showing the spatial relationship between the high Cu-As sulphide domain (cluster 3; green drill strings) and the low $\mathbf{C u}$ - high Zn sulphide domain (cluster 1; purple drill strings).

Cluster 5 corresponds to a potassium and aluminium rich domain that slightly overprints the vuggy silica alteration on the margin of the deposit. This domain is characterized by high sulphate content consistent with the appearance of the sulphate minerals alunite and jarosite (Figure 14). This domain extends beyond the logged vuggy silica zone into the advanced argillic alteration halo within the host rock (see figure 15). Due to the presence of sulphate minerals, the ore in this domain has a higher lime consumption and is slightly softer than the vuggy silica ore. Characterization of this domain is relevant for mine planning because these variables can affect processing costs and throughput estimates.


Figure 14 Drill core photo of interval alunite rich domain (cluster 5). These rocks are easily identified by their distinct yellow-beige colour


Figure 15 Plan view of the logged vuggy silica zone (grey) and logged advanced argillic alteration (yellow) showing the spatial relationship between the alunite-rich domain (light blue drill strings) and the logged geology and alteration.

## 4 Conclusions

Geochemical domains were created using a $k$-means clustering algorithm to analyze multi-element data from the vuggy silica zone of a high sulphidation epithermal gold deposit (Deposit A). This analysis was able to partition the deposit into several distinct geochemical domains that correlate with the known geological and alteration domains of the deposit, but also provided additional information on the delineation of the transition zone between the oxide and sulphide zones.

The transition zone at the oxide-sulphide interface includes secondary chalcocite, $\mathrm{Cu}_{2} \mathrm{~S}$, which is soluble in cyanide solutions. When cyanide consumption and copper gold recoveries were examined the transition domain (cluster 2) was found to contain higher soluble copper concentrations than the oxide domain (cluster 4). Gold recoveries within the transition domain (cluster 2) were on the average of $75 \%$. Within the oxide domain (cluster 4) gold recoveries were approximately $83 \%$.

Because the gold recoveries are so variable between the oxide and transition zone for Deposit A the robust, unbiased delineation of the oxide and transition zones is critical to assessing the economic viability of this project. Compositional data analysis, including robust principal component analysis and $k$-means cluster analysis is an effective tool for increasing the understanding and knowledge of an orebody.

## References

Aitchison, J., (1986), The statistical analysis of compositional data. Monographs on statistics and applied probability. Chapman \& Hall Ltd., London

Egozcue, J. J., et al., (2003) Isometric log ratio transformations for compositional data analysis. Mathematical Geology. 35(3):279-300

Filmozer, P., and Hron, K., (2011), "Robust statistical analysis". In V. Pawlowsky-Glahn and A. Buccianti, editors, Compositional Data Analysis. Theory and Applications, pp. 59-71, John Wiley \& Sons, Chichester (UK)

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

Templ, M., K. Hron, and P. Filzmoser (2011). robCompositions: Robust Estimation for Compositional Data. Manual and package, version 1.9.0 http://cran.rproject.org/web/packages/robCompositions/index.html

Hennig, C. (2014). fpc: Flexible procedures for clustering. R package version 2.1-9 http://CRAN.R-project.org/package=fpc

Palarea-Albaladejo, J., and J. A., Martin-Fernandez (2014). zCompositions: Imputation of Zeros and Nondetects in Compositional Data Sets. R package version 1.0.3. http://CRAN.Rproject.org/package=zCompositions

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

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

# Regression model for compositional data using a spatial econometric approach 

T. Yoshida ${ }^{1}$ and M. Tsutsumi ${ }^{2}$<br>${ }^{1}$ Graduate School of Systems and Information Engineering, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan; yoshida.takahiro@sk.tsukuba.ac.jp<br>${ }^{2}$ Faculty of Engineering, Information and Systems, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8573, Japan;<br>tsutsumi@sk.tsukuba.ac.jp


#### Abstract

In recent years, regression models for compositional data that consider spatial dependency have been discussed mainly in the field of geology. In compositional data analysis, most methods integrating spatial dependency have been geo-statistical approaches; research using spatial econometric approaches has been limited to a few recent studies. Furthermore, these studies' methods for considering spatial dependency have not taken into account the effects of neighboring pixels, commonly used in the field of image processing. Thus, the present study extends these studies' approaches to spatial dependency by presenting an empirical regression model for compositional data with a spatial econometric approach. To do so, we briefly review past work on relevant regression model, and then, propose a method under which spatial dependency is assumed to decrease with geographical distance. We then test the method through an empirical examination using land-use data with the characteristics of compositional data; this allows us to evaluate the proposed method in terms of predictive accuracy.


Key words: compositional data, spatial autocorrelation, spatial econometrics, landuse data.

## 1 Introduction

Research and empirical study of the statistical model of compositional data have been developed thus far as a methodology of geology. Thus, they are biased toward the geological field via, for example, composition investigation of mineral resources. On the other hand, there is less accumulated knowledge and empirical studies with regard to socioeconomic data than geological data.

In recent years, there has been increasing need to consider the spatial effects typified by positive spatial autocorrelation, which can be expressed as "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). When analyzing data with location information, the use of spatial information of the data improves the estimation and prediction accuracy of the statistical model and it is possible to increase reliability.

Consideration of spatial effects on the analysis of compositional data has been triggered by geological requirements for deep-mine studies, although there are many other examples of using the technique of geo-statistics to handle data as a continuous amount of space, which involves separate counting units. However, to the best of our knowledge, there are only limited studies using the technique of spatial econometrics dealing with data as a discrete amount of space.

Based on the current situation, this study aims to undertake an analysis using socioeconomic data that has never been covered sufficiently by compositional data model. From the viewpoint of applicability to compositional data model to socioeconomic data, this attempt is significant for the promotion of accumulated research. Moreover, by applying the spatial effects to compositional data model, we can consider even socioeconomic data using the technique of spatial econometrics, which indicates the usefulness of the study in comparing different setting of spatial effects.

## 2 Statistical model of compositional data

### 2.1 Log ratio transformation

The compositional data for the sample $\mathbf{y}_{i}=\left(y_{i 1}, \ldots, y_{i D}\right)$ in a $D$-dimension refer to data that have a constant sum constraint (Equation (1)):

$$
\begin{equation*}
y_{i 1}+\cdots+y_{i D}=1, y_{i 1}, \ldots, y_{i D}>0 \tag{1}
\end{equation*}
$$

Here, $n$ is the number of samples. This is a compositional data condition because it belongs to simplex space that is included in real space where it is inappropriate to apply compositional data techniques of statistical analysis directly (Aitchison, 1986). Use of a statistical analysis technique that is defined in the real space of the compositional data, is necessary to map the compositional data from a simplex space into real space, and such a technique is employed often as a method of log ratio transformation.

Log ratio transformation is divided by common standardized variables in each attribute of the compositional data, and replaces the value for each attribute in the ratio of the transformation operation by taking the logarithm to the ratio (Aitchison, 1986). Here, this study takes the ratio to solve a constant sum constraint and the ratio between the compositional data is used to approximate a log-normal distribution closely. In addition, this operation can approximate the compositional data after transformation into easily normal distribution analysis by taking the logarithm.

This study directly uses a dimension of compositional data in standardized variables, using an additive $\log$ ratio transformation, which is a typical log ratio transformation method (Equation (2)):

$$
\begin{equation*}
z_{i k}=\ln \left(y_{i k} / y_{i D}\right), \quad k=1, \ldots, d . \tag{2}
\end{equation*}
$$

Here, $y_{i D}$ are normalized variables, $d=D-1$.

### 2.2 Multivariate conditional autoregressive model

In this study, in line with Allen and others (2013), we can consider spatial autocorrelation by incorporating spatial random effects into the multivariate conditional autoregressive (MCAR) model proposed by Mardia (1988).
Now, in Equation (3), the dependent variable $\mathbf{z}_{i}=\left(z_{i 1}, \ldots, z_{i d}\right)$ is assumed to follow $d$-dimensional normal distribution $N_{d}$ :

$$
\begin{equation*}
\mathbf{z}_{i} \sim N_{d}\left(\mathbf{B}^{T} \mathbf{x}_{i}+\boldsymbol{\eta}_{i}, \mathbf{V}\right) \tag{3}
\end{equation*}
$$

Here, $\mathbf{B}$ is a parameter matrix of $(p+1) \times d, \mathbf{x}_{i}$ is an explanatory variable vector of $(p+1) \times 1, \boldsymbol{\eta}_{i}$ is a random effect vector of $d \times 1$, and $\mathbf{V}$ is the variance-covariance matrix of $d \times d, T$ is the symbol representing the transpose. Moreover, in order to consider spatial autocorrelation, as Equation (4) shows, prior distribution of the autoregressive type $\boldsymbol{\eta}_{\boldsymbol{i}}$ is assumed:

$$
\begin{equation*}
\boldsymbol{\eta}_{i} \left\lvert\,\left\{\boldsymbol{\eta}_{j}\right\}_{j \neq i} \sim N_{d}\left(\frac{1}{S_{i}} \sum_{j=1}^{n} w_{i j} \boldsymbol{\eta}_{j}, \frac{1}{S_{i}} \boldsymbol{\Sigma}\right)\right. \tag{4}
\end{equation*}
$$

Here, $S_{i}$ is the row sum of the $i$-th row of the spatial weight matrix $\mathbf{W}$ (described in section 3) of $n \times n$ with $w_{i j}$ to the elements, $\boldsymbol{\Sigma}$ is the variance-covariance matrix of $d \times d$.

### 2.3 Parameter estimation method

Since the parameter estimation method of the MCAR model is used widely in Bayesian estimation with the Markov chain Monte Carlo (MCMC) method, we use this estimation method in this study.

Now, with Bayesian estimation, prior distribution of the parameters $\mathbf{B}, \mathbf{V}$, and $\boldsymbol{\Sigma}$ are assumed to be independent and are set as the respective prior distribution Equations (5)-(7):

$$
\begin{gather*}
\operatorname{vec}(\mathbf{B}) \sim N_{d(p+1)}\left(\operatorname{vec}\left[\boldsymbol{\mu} \mathbf{0}_{d \times p}\right], \lambda \mathbf{1}_{d(p+1)}\right),  \tag{5}\\
\mathbf{V} \sim I W_{d}\left(m_{V}, \mathbf{M}_{V}\right)  \tag{6}\\
\boldsymbol{\Sigma} \sim I W_{d}\left(m_{\Sigma}, \mathbf{M}_{\Sigma}\right) \tag{7}
\end{gather*}
$$

Here, $I W_{d}$ is the $d$-dimensional inverse Wishart distribution, vec () is an operator to convert the matrix to a column vector, $\boldsymbol{\mu}=\left(\mu_{1}, \ldots, \mu_{d}\right)$ is a vector of $d \times 1$ with $\mu_{k}=\mathbf{x}_{i}^{T} \boldsymbol{\beta}_{k}$ to the elements, $\boldsymbol{\beta}_{k}$ is $(p+1) \times 1$ of the parameter vector $k$ column of $\mathbf{B}, \mathbf{0}_{d \times p}$ is the matrix of $d \times p$ with 0 to all the elements, the matrix of $\mathbf{1}_{d(p+1)}$ is $d(p+1) \times d(p+1)$ with 1 to all the elements, and $\lambda, m_{V}$, $\mathbf{M}_{V}, m_{\Sigma}$, and $\mathbf{M}_{\Sigma}$ are hyperparameters, which will be described later.

From Bayes' theorem, the joint probability density function expressed by Equations (3) and (4), and the prior distribution expressed by Equations (5)-(7), can be derived by the conditional posterior distribution of Equations (8)-(11):

$$
\begin{gather*}
v e c(\mathbf{B}) \mid \mathbf{z}_{i}, \mathbf{X}, \mathbf{V}, \boldsymbol{\eta}_{i} \sim N_{1}\left(\boldsymbol{\Omega}^{-1}\left(\mathbf{V}^{-1} \otimes \mathbf{X}^{T} \mathbf{X}\right)\left(\mathbf{X}^{T} \mathbf{X}\right)^{-1} \sum_{i=1}^{n} \mathbf{x}_{i}\left(\mathbf{z}_{i}-\boldsymbol{\eta}_{i}\right), \boldsymbol{\Omega}^{-1}\right)  \tag{8}\\
\mathbf{V} \mid \mathbf{z}_{i}, \mathbf{X}, \mathbf{B}, \boldsymbol{\eta}_{i} \sim I W_{d}\left(m_{V}+S_{i}, \mathbf{M}_{V}+\sum_{i=1}^{n} \mathbf{E}_{i} \mathbf{E}_{i}^{T}\right)  \tag{9}\\
\boldsymbol{\eta}_{i} \mid \mathbf{z}_{i}, \mathbf{X}, \mathbf{B},\left\{\boldsymbol{\eta}_{i}\right\}_{j \neq i}, \mathbf{\Sigma} \sim N_{d}\left(\mathbf{A}^{-1}\left(\mathbf{V}^{-1}\left(\mathbf{z}_{i}-\mathbf{B}^{T} \mathbf{x}_{i}\right)+\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n} w_{i j} \boldsymbol{\eta}_{j}\right)\right), \mathbf{A}^{-1}\right),  \tag{10}\\
\boldsymbol{\Sigma} \mid \boldsymbol{\eta}_{i} \sim I W_{d}\left(m_{\Sigma}+S_{i}, \mathbf{M}_{\Sigma}+\sum_{i=1}^{n} \sum_{j=1}^{n}\left(\mathbf{D}_{W}-\mathbf{W}\right)_{i j} \boldsymbol{\eta}_{i} \boldsymbol{\eta}_{i}^{T}\right) \tag{11}
\end{gather*}
$$

Here, $\mathbf{X}$ is the explanatory variable matrix of $n \times(p+1), \boldsymbol{\Omega}=\lambda \mathbf{I}_{d(p+1)}+\mathbf{V}^{-1} \otimes \mathbf{X}^{T} \mathbf{X}, \mathbf{I}_{d(p+1)}$ is the identity matrix of $d(p+1) \times d(p+1), \mathbf{E}=\mathbf{z}_{i}-\mathbf{B}^{T} \mathbf{x}_{i}-\boldsymbol{\eta}_{i}, \mathbf{A}=\mathbf{V}^{-1}+S_{i} \boldsymbol{\Sigma}^{-1}, \mathbf{D}_{W}$ is a diagonal matrix of $n \times n$ with $S_{i}$ to the diagonal elements, and $\otimes$ is the Kronecker product. From Equations (8)-(11), the conditional posterior distribution of parameters for are both standard distribution, thereby enabling efficient parameter estimation by the MCMC method using Gibbs sampling.

## 3 Spatial autocorrelation in the model

In this study, in order to take into account the spatial autocorrelation of compositional data in the statistical model, previous research (e.g., Allen and others, 2013) of the simple Rook type used in $\mathbf{W}$ (WRook) (Figure 1A) rather than more socioeconomic data of the Queen type is believed to represent spatial characteristics with $\mathbf{W}$ (WQueen) (Figure 1B). Much of the socioeconomic data are spatial data with geographic information, which has the property of Tobler (1970). When dealing with spatial data, it is desirable to assume that $\mathbf{W}$ better represents the characteristics.


Figure 1: Comparison of how to set the degree of spatial autocorrelation (elements of $\mathbf{W}$ ) to the center of the mesh

This study attempts to extend the nature of the relationships of spatial autocorrelation (degree) in close proximity (Figures 1 A and 1 B ) in the future, based on the distance considered to be the standard, in order to apply the technique of spatial econometrics accumulated in the methodology of using $\mathbf{W}$ in the statistical model of compositional data.

We use two models in which $\mathbf{W}$ differs from that described in Section 3; specifically, we compare prediction accuracy by only the difference in the setting of the spatial weight matrix $\mathbf{W}$. We call these the SPRook and SPQueen models, using the WRook and WQueen spatial weight matrix $\mathbf{W}$, respectively.

## 4 Empirical analysis

### 4.1 Data overview

In this study, we use land-use data from Ibaraki Prefecture in Japan, $(n=5,904)$ as the compositional data. Here, we describe the data used. The dependent variable is digital national land information that is available from land-use third-order mesh data for fiscal 2006 in Ibaraki Prefecture. The data for each mesh is stored as the land area occupied. If calculating the ratio, the data can be regarded as compositional data. Furthermore, since a zero value is not allowed in the ratio of each land use in Equation (2), this study follows Allen and others (2013) in using a minute value ( 0.01 ) to calculate the ratio in the terms added to each land use. The explanatory variables are 13 variables as follows: the logarithm of population density [people $/ \mathrm{m} 2$ ], average elevation $[\mathrm{m}$ ], average inclination [degrees], total road length [km] distance to the nearest railway from the mesh center of gravity $[\mathrm{km}]$, distance to the nearest first-grade river $[\mathrm{km}]$, a microtopography (alluvial fan, natural levee, marsh, delta, state sandbar-gravel) dummies, a lakes dummy, and an urbanization control area dummy.

### 4.2 Prediction results

For the prediction, in the normalized variable $y_{i D}$ of Equation (2), we adopted the fewest categories of composition in land-use, and for the initial value of the hyperparameters, $\lambda=1,000, m_{V}=$ $m_{\Sigma}=(d+2)$, we set $\mathbf{M}_{V}=\mathbf{M}_{\Sigma}=2 \mathbf{I}_{d}$. There are 20,000 random numbers generated in the Gibbs sampling and the burn-in period is 2,000 times.

The evaluation of the prediction accuracy used is the Aitchison distance (Equation (12))

$$
\begin{equation*}
A D_{i}\left(\mathbf{y}_{i}, \hat{\mathbf{y}}_{i}\right)=\sqrt{\sum_{K=1}^{D}\left[\ln \left\{\frac{y_{i K}}{\left(\prod_{K=1}^{D} y_{i K}\right)^{1 / D}}\right\}-\ln \left\{\frac{\hat{y}_{i K}}{\left(\prod_{K=1}^{D} \hat{y}_{i K}\right)^{1 / D}}\right\}\right]^{2}} \tag{12}
\end{equation*}
$$

which represents the similarity between the compositional data. Table 1 summarizes the basic statistics for the predicted and measured values of the two models to calculate the respective targets. Incidentally, the predicted values of the compositional data are obtained by performing inverse transformation of Equation (2) to the posterior mean of the predicted values of $\mathbf{z}_{i}$ (Aitchison, 1986). This is because it does not follow the normal distribution performed in a Wilcoxon signed-rank test. This results in $p=0.00(<0.01)$, supporting a difference in the median SPRook and SPQueen (Median) models at the $1 \%$ significance level. That is, there is a significant statistical improvement in the prediction accuracy for the Queen-type spatial weight matrix over the Rook type.

Table 1: Descriptive statistics of $A D_{i}\left(\mathbf{y}_{i}, \hat{\mathbf{y}}_{i}\right)$

| Model | Min. | Median | Mean | Max. | SD |
| :--- | ---: | ---: | ---: | ---: | ---: |
| SPRook | 0.010 | 0.292 | 0.457 | 2.737 | 0.528 |
| SPQueen | 0.009 | 0.263 | 0.424 | 2.683 | 0.533 |

## 5 Concluding remarks

This study undertook a prediction of compositional data by using a spatial weight matrix representing proximity. It found there is higher expressive power than the spatial weight matrix
representing adjacency used in previous studies. This result suggests that a method based on distance is effective for a spatial weight matrix of the statistical model of socioeconomic compositional data.

## References

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. London: Chapman and Hall.

Allen, J., Leininger, T., Hurd, J., Civco, D., Gelfand, A., and Silander, J. (2013). Socioeconomics drive woody invasive plant richness in New England, USA through forest fragmentation. Landscape Ecology 28(9), pp. 1671-1686.

Mardia, V. (1988). Multi-dimensional multivariate Gaussian Markov random fields with applications to image processing. Journal of Multivariate Analysis 24(2), pp. 265-284.

Tobler, W. (1970). A computer movie simulating urban growth in the Detroit region. Economic Geography 46 (2), pp. 234-240.

# Application of compositional data analysis in biomonitoring of atmospheric dust precipitation 

Z. Ziembik, A. Dołhańczuk-Sródka<br>Independent Chair of Biotechnology and Molecular Biology, Opole University, Poland; ziembik@uni.opole.pl


#### Abstract

Biomonitoring is a method of environmental pollution estimation. It involves the use of living organisms to assess type and degree of environment contamination. One of the species most frequently used in biomonitoring is moss. It accumulates substances deposited from air, among them the dust containing different contaminants. Composition of a precipitate is determined both by distant emission sources and local secondary redeposition. It is highly desirable to differentiate between both deposition components. Distinction between them would be helpful in spatiotemporal characterization of pollution and estimation of its sources. Combustion of fossil fuels and some industrial activities produce large amounts of dust containing radioactive isotopes, which are then deposited on the ground surface. These include natural radioisotopes and their decay products as well as some artificial radioisotopes. In our studies samples of moss and its ground base were collected in Bory Stobrawskie forest, located in southwestern part of Poland. In samples specific activity concentrations of gammaradioactive isotopes Cs-137, K-40, Bi-214 and $\mathrm{Pb}-210$ were determined. The specific activity concentrations were recalculated to mass concentrations. In the results exploration the methods designed for the compositional data analysis were used. In cluster analysis a matrix of distances between points was taken for dissimilarity matrix. To study relationship between the radioisotope concentrations a method based on actual data resampling was proposed.


Key words: biomonitoring, soil, moss, radioisotope, covariability.

## Introduction

In many production branches, as a result of industrial activities, large amounts of dust containing radioactive isotopes is produced. These include uranium, radium, thorium and their decay products, which together with the produced dust are spread over large areas (Borrego et al., 2007; Bou-Rabee et al., 2009; Peroni et al., 2012; Abdel Rahman et al., 2013). Another source of radionuclides in the environment are dusts arising after the nuclear test explosions carried out in large numbers especially in the late 50 s and 60 s of the $20^{\text {th }}$ century. These dust falling gradually covered the entire surface of the Earth (Simon, Bouville, and Beck 2004; Tsumune et al., 2011). Though their radioactivity is often low and does not create threats to health of living organisms, they could be used in estimation of other contaminants spread in environment.

The dust comprising radioisotopes can be transported in atmosphere in the long distances. When the dust is deposited on surface of earth it becomes an element of local environment. Discrimination between the fallen dust and the local matter can help in localization of pollution sources. An information about dispersion range of pollutants could be also achieved in such way.

The aim of the studies was investigation of radioisotopes content in moss and in surface soil layer. It was expected that relationships between different radioisotopes content in plant and soil matrix will provide information about the deposited dust sources.

## Sampling locations and measurements

Samples of the mosses (Pleurozium schreberi) and soil surface layer, were taken at 21 places in forests around Opole (south-western Poland) in an area of Bory Stobrawskie limited approximately by a $40 \mathrm{~km} \times$ 20 km rectangle. The moss samples were taken manually from forest glades at the places located not less than 200 m from roads. Subsamples were kept at room temperature for a few days and then dried at $105^{\circ} \mathrm{C}$ until constant mass was reached (usually about 24 h ). Before measurement, subsamples from each sampling place were carefully mixed together and a portion of the mixture was placed in a Marinelli container.

The measurement of radionuclide activity in moss and soil samples was carried out by means of a gamma-spectrometer with a germanium detector HPGe (Canberra) of high resolution: 1.29 keV (FWHM) at 662 keV and 1.70 keV (FWHM) at 1332 keV . Relative efficiency: $21.7 \%$. Energy and efficiency calibration of the gamma spectrometer was performed with the standard solutions type MBSS 2 (Czech Metrological Institute, Praha). The geometry of the calibration source was a Marinelli container $\left(447.7 \pm 4,5 \mathrm{~cm}^{3}\right)$, Time of measurement was 24 h for all samples. Measuring process and analysis of spectra were computer controlled with use of the software GENIE 2000. The results were corrected to the same date of measurement.

## Computations

In data analysis, clustering methods were used. Results of unsupervised clustering method application provided assessment of structures presence in data. Initial results supposing number of clusters were used in fuzzy clustering analysis. For each sampling point contribution of a cluster was calculated. To estimate contribution of the $k$-th radioisotope in the $j$-th cluster structure, value of the following expression can be considered [Eq. 1]:

$$
\begin{equation*}
F C_{j}=\sum_{i=1}^{n} f_{j i} c_{k i} \tag{1}
\end{equation*}
$$

where $f_{\mathrm{ji}}$ is contribution of the j -th cluster at sampling site $i$ and $c_{\mathrm{ij}}$ is fraction of the $k$-th isotope at the same site.

Matrix representation FC of Equation 1 can be formulated as follows [Eq. 2]:

$$
\begin{equation*}
\mathbf{F C}=\mathbf{F}^{\prime} \cdot \mathbf{C} \tag{2}
\end{equation*}
$$

where $\mathbf{F}$ is cluster contribution matrix, $\mathbf{C}$ is concentration (fraction) matrix and ' is transposition operation.

Covariabilty character of two components can be deduced from variability of the components ratio, defined in the relationship [Eq. 3]:

$$
\begin{equation*}
V R=\operatorname{var}\left(\log \frac{\boldsymbol{c}_{\mathrm{A}}}{\boldsymbol{c}_{\mathrm{B}}}\right) \tag{3}
\end{equation*}
$$

Increase in both components values results in constant (for proportionality) or almost constant logratio of these variables. In opposite, decrease in one component value with increase in values of the second one produces differentiated logratios. In pairs sorted, for example, by increasing values of the first variable A, the ratio of the first pairs would contain a small number in nominator and a big number in denominator. The corresponding ratio would be small. In ratios calculated from the pairs close to the end of the sorted data, the number in nominator would be big and the one in nominator would be small. The logratios would be much bigger than the ones calculated from the pairs at the sorted data beginning. The relationship in which local extreme values are observed, are described by intermediate logratio variabilities. Similar effect can be expected for independent variables.

Assessment of covariability type on the base of logratio variance is not obvious. It is difficult to deduce covariability type from raw logratio variance without information about its range of changes. Though the lower limit of the variance is 0 , its upper limit is not defined. To solve this problem resampling method can be used. From the two actual data vectors, describing content of two components in a system, random samples of data can be drawn. For each drawing variance of logratio can be calculated. In this way conditional distribution of the variance can be estimated.

Comparison of the actual $V R$ value with its distribution obtained from the data resampling allow to estimate type of the variables covariability. If the actual $V R$ value is low among the others calculated from the resampled data, then an increase in A variable with increase in B can be supposed. When the actual $V R$ value is high among the other, then an increase in A variable with decrease in B can be concluded.

All statistical computations were performed using R language ( $R$ Development Core Team 2015). In cluster analysis the R package "cluster" was used (Maechler et al., 2014; Kaufman and Rousseeuw 2005).

## Results and discussions

Activity concentrations of Cs-137, K-40, Bi-214 and $\mathrm{Pb}-210$ were determined in moss and soil samples. The Cs-137 is artificial isotope with half-life time 30.1 years. Its main source in environment are the nuclear test explosions carried out in large numbers especially in the late 50 s and 60 s of the $20^{\text {th }}$ century and accidents in nuclear power plants (Chernobyl 1986, Fukushima Da-Ichi 2011) (Aleksakhin, Sanzharova, and Fesenko 2006; UNSCEAR 2008; Hirose 2012; Akahane et al., 2012). The radioactive dust falling gradually covered the entire surface of the Earth.

The long-living K-40 radioisotope with half-life time $1.25 \cdot 10^{9}$ years, is common in nature. It is strictly bounded with natural potassium in which its abundance is $0.0119 \%$. The Bi-214 (half-life time 19.9 min ) and $\mathrm{Pb}-210$ (half-life time 22.2 years) are members of the radium decay chain.

Activity concentrations of the radioisotopes were recalculated to mass fractions from the formula [Eq. 4]:

$$
\begin{equation*}
c=\frac{t_{1 / 2} M}{\ln 2 N_{\mathrm{A}}} a \tag{4}
\end{equation*}
$$

where $c$ is mass fraction of the radionuclide, $t_{1 / 2}$ is its half-life time, $M$ is its molar mass, $N_{\mathrm{A}}$ is Avogadro constant and $a$ is specific activity concentration of a radionuclide.

In Table 1 the statistical summary of the calculated concentrations (fractions) is shown. The min and max are, respectively, minimum and maximum values, $q_{25}$ and $q_{75}$ are lower and upper quartiles, $q_{50}$ is median, mean is arithmetic mean, $S D$ is standard deviation and $M A D N$ is normalized median absolute deviation about the median (Maronna, Martin, and Yohai 2006).

Table 1. Statistical parameters of the radioisotopes concentrations in moss and in soil

|  | moss |  |  |  |  |  | soil |  |  |  |
| :--- | ---: | ---: | ---: | ---: | ---: | ---: | ---: | :---: | :---: | :---: |
|  | Cs-137 | $\mathrm{K}-40$ | $\mathrm{Bi}-214$ | $\mathrm{~Pb}-210$ | $\mathrm{Cs}-137$ | $\mathrm{~K}-40$ | $\mathrm{Bi}-214$ | $\mathrm{~Pb}-210$ |  |  |
|  | $\cdot 10^{13}$ | $\cdot 10^{7}$ | $\cdot 10^{20}$ | $\cdot 10^{14}$ | $\cdot 10^{13}$ | $\cdot 10^{7}$ | $\cdot 10^{20}$ | $\cdot 10^{14}$ |  |  |
| min | 0.07 | 5.48 | 0.08 | 4.12 | 0.03 | 4.09 | 0.22 | 0.66 |  |  |
| $q_{25}$ | 0.47 | 7.20 | 0.19 | 5.76 | 0.08 | 4.95 | 0.33 | 1.21 |  |  |
| $q_{50}$ | 1.19 | 8.46 | 0.21 | 10.35 | 0.25 | 6.08 | 0.40 | 2.19 |  |  |
| mean | 1.30 | 8.20 | 0.94 | 9.89 | 0.60 | 5.95 | 0.40 | 2.36 |  |  |
| $q_{75}$ | 1.52 | 8.92 | 1.13 | 13.95 | 0.95 | 6.65 | 0.45 | 2.91 |  |  |
| max | 3.84 | 11.20 | 3.66 | 15.33 | 2.17 | 8.49 | 0.58 | 6.10 |  |  |
| SD | 1.27 | 1.84 | 1.29 | 4.68 | 0.67 | 1.22 | 0.10 | 1.57 |  |  |
| MADN | 0.60 | 0.94 | 0.20 | 6.63 | 0.33 | 1.12 | 0.09 | 1.54 |  |  |

An assessment of the structure presence in data was performed. To estimate groups formation in data, distances between compositional points A and B in simplex space $d_{\mathrm{a}}$ were calculated, using Equation 5:

$$
\begin{equation*}
d_{\mathrm{a}}=\sum_{j=1}^{D} \ln \frac{C_{A j}}{\mathrm{~g}\left(\mathbf{c}_{A}\right)} \ln \frac{\mathrm{c}_{B j}}{\mathrm{~g}\left(\mathbf{c}_{B}\right)} \tag{5}
\end{equation*}
$$

where $\mathrm{g}(\mathbf{c})$ is geometric mean of components concentrations.
The distances matrix was used in construction of dendrograms. Both divisive and agglomerative clustering methods were used. Additionally each branch in dendrogram was marked by symbol of the material type, "s" for soil and " $m$ " for moss. Structures of the dendrograms are shown in Figure 1. In both of them two groups of data can be distinguished, but the material type is not crucial in the separation observed.

In further analysis fuzzy clustering method was used. In description of grouping trends, existence of two clusters were assumed. From cluster contributions and concentrations, elements of the FC matrix were calculated. In Table 2 the computation result is shown. Significant difference in $F C$ values for Cs-137 was observed in clusters. It is a result of uneven contamination of the studied region with Cs-137 after Chernobyl NPP breakdown. It can be supposed that the observed clustering is not related to spatial distribution of other than Cs-137 radioisotopes determined.

Covariability of a radioisotope concentrations in moss and soil was tested using $V R$ coefficient. On the base of $V R$ conditional distribution, estimated from the actual data resampling, the type of reciprocal changes in radionuclides concentrations was studied.

As an example, the results of studies on the relationship between $\mathrm{Cs}-137$ and $\mathrm{Pb}-210$ in moss are described. In Figure 2 graphs of the empirical probability distribution function (EPDF) a) and the empirical cumulated distribution function (ECDF) b) are shown. Actual, calculated from the data, $V R_{\mathrm{a}}$ is marked with vertical, dashed line. The $V R$ distribution is unimodal, and particularly low parameter values correspond to positive covariability (low ECDF values) and particularly high values correspond to negative covariability (high ECDF values). Intermediate $V R$ describes significantly non-linear or multimodal relationships, or mutually independent variables.

In the example position of $V R_{\mathrm{a}}$ in graphs does not confirm linear dependence between $\mathrm{Cs}-137$ and $\mathrm{Pb}-210$ concentrations in population, neither positive nor negative.

The $V R$ estimated distribution function enables the quantiles calculation. In this way hypotheses concerning $V R$ value, and covariability character, can be tested. In Table 3 the actual $V R_{\mathrm{a}}$ values (upper part of cell) and ranges of the $95 \%$ confidence intervals (lower part of cells) are shown for relationships between radioisotopes concentrations in moss and in soil.


Figure 1: Structures of the distances dendrogram obtained using agglomerative a) and divisive b) methods.

Table 2. Elements of the FC matrix calculated for 2 clusters

| $F C$ | $\mathrm{Cs}-137$ | $\mathrm{~K}-40$ | $\mathrm{Bi}-214$ | $\mathrm{~Pb}-210$ |
| :---: | :---: | :---: | :---: | :---: |
| Cluster 1 | 0.6 | 8.1 | 6.4 | 4.9 |
| Cluster 2 | 1.4 | 9.0 | 7.7 | 6.5 |

The results shown in Table 3 suppose absence of reliable linear relationships between radioisotopes concentrations in moss and in soil.

Table 3. The $V R$ values and ranges of $95 \%$ confidence intervals estimated from actual data resampling

|  | radioisotopes | Cs-137 | K-40 | Bi-214 |
| :---: | :---: | :---: | :---: | :---: |
| moss | K-40 | $\begin{gathered} 1.6 \\ 1.4-1.9 \end{gathered}$ |  |  |
|  | Bi-214 | $\begin{gathered} 3.9 \\ 1.7-7.7 \end{gathered}$ | $\begin{gathered} 1.6 \\ 1.6-4.8 \end{gathered}$ |  |
|  | $\mathrm{Pb}-210$ | $\begin{gathered} 2.4 \\ 1.4-2.9 \end{gathered}$ | $\begin{gathered} 0.39 \\ 0.29-0.57 \end{gathered}$ | $\begin{gathered} 1.2 \\ 1.0-3.4 \end{gathered}$ |
| soil | K-40 | $\begin{gathered} 1.7 \\ 1.3-2.4 \end{gathered}$ |  |  |
|  | Bi-214 | $\begin{gathered} 1.8 \\ 1.7-2.4 \end{gathered}$ | $\begin{gathered} 0.053 \\ 0.061-0.15 \end{gathered}$ |  |
|  | $\mathrm{Pb}-210$ | $\begin{gathered} 1.4 \\ 1.6-3.2 \end{gathered}$ | $\begin{gathered} 0.50 \\ 0.36-0.61 \end{gathered}$ | $\begin{gathered} 0.42 \\ 0.34-0.64 \end{gathered}$ |

The $V R_{\mathrm{a}}$ value for $\mathrm{K}-40$ and $\mathrm{Bi}-214$ in moss and soil, as well as $\mathrm{Pb}-210$ and $\mathrm{Cs}-137$ in soil suppose linear relationship between corresponding concentrations.


Figure 2: Graphs of the empirical probability distribution function (EPDF) a) and the empirical cumulated distribution function (ECDF) b) of reciprocal relationship between $\mathrm{Cs}-137$ and $\mathrm{Pb}-210$ concentrations.

Low values of $V R$ suppose proportionality of the listed radioisotopes concentrations, though they are different in moss and in soil. No negative covariabilities (increase-decrease) in concentrations were observed.

## Conclusions

The proposed approach to compositional data analysis provides methods of covariabilties estimation between a system components. Statistical inference based on covariance analysis can be replaced by the VR consideration. In compositional data mining the distances between points in simplex space can be used in structures recognition.

## Acknowledgments

The Project received financial assistance from the funds of the National Science Centre, granted by force of the decision no. DEC-2011/03/D/ST10/05392.

## References

Abdel Rahman, R.O., M. Elmesawy, I. Ashour, and Y.-T. Hung. 2013. „Remediation of NORM and TENORM Contaminated Sites-Review Article".
Akahane, Keiichi, Shunsuke Yonai, Shigekazu Fukuda, Nobuyuki Miyahara, Hiroshi Yasuda, Kazuki Iwaoka, Masaki Matsumoto, Akifumi Fukumura, and Makoto Akashi. 2012. „The Fukushima Nuclear Power Plant accident and exposures in the environment". The Environmentalist 32 (2): 136-43. doi:10.1007/s10669-011-9381-2.
Aleksakhin, R.M., N.I. Sanzharova, and S.V. Fesenko. 2006. „Radioecology and the accident at the Chernobyl nuclear power plant". Atomic Energy 100 (4): 257-63.

Borrego, E., J.L. Mas, J.E. Martín, J.P. Bolívar, F. Vaca, and J.L. Aguado. 2007. „Radioactivity levels in aerosol particles surrounding a large TENORM waste repository after application of preliminary restoration work". Science of The Total Environment 377 (1): 27-35. doi:10.1016/j.scitotenv.2007.01.098.
Bou-Rabee, F., A.Z. Al-Zamel, R.A. Al-Fares, and H. Bem. 2009. „Technologically Enhanced Naturally Occurring Radioactive Materials in the Oil Industry (TENORM). A Review". Nukleonika 54 (1): 3-9.
Hirose, Katsumi. 2012. „2011 Fukushima Dai-ichi nuclear power plant accident: summary of regional radioactive deposition monitoring results". Journal of Environmental Radioactivity 111: 13-17. doi:10.1016/j.jenvrad.2011.09.003.
Kaufman, L., and P. J. Rousseeuw. 2005. Finding groups in data. An introduction to cluster analysis. New York: Wiley.
Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, and K. Hornik. 2014. „cluster: Cluster Analysis Basics and Extensions. R package version 1.15.3".
Maronna, Ricardo A, R. Douglas Martin, and Víctor J Yohai. 2006. Robust Statistics: Theory and Methods. Chichester, England: J. Wiley.
Peroni, M., V. Mulas, E. Betti, L. Patata, and P. Ambrosini. 2012. „Decommissioning and Remediation of NORM/TENORM Contaminated Sites in Oil and Gas". Chemical Engineering Transactions 28: 181-86. doi:10.3303/CET1228031.
R Development Core Team. 2015. R: A language and environment for statistical computing. $R$ Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.
Simon, S.L., A. Bouville, and H.L. Beck. 2004. „The geographic distribution of radionuclide deposition across the continental US from atmospheric nuclear testing". Journal of Environmental Radioactivity 74 (1-3): 91-105.
Tsumune, Daisuke, Michio Aoyama, Katsumi Hirose, Frank O. Bryan, Keith Lindsay, and Gokhan Danabasoglu. 2011. „Transport of 137Cs to the Southern Hemisphere in an ocean general circulation model". Progress In Oceanography 89 (1-4): 38-48. doi:16/j.pocean.2010.12.006.
United Nations. 2008. „UNSCEAR 2008 Report Vol. I. Sources of ionizing radiation. United Nations Scientific Committee on the Effects of Atomic Radiation UNSCEAR 2008 Report to the General Assembly, with scientific annexes. AnnexB". http://www.unscear.org/docs/reports/2008/1180076_Report_2008_Annex_D.pdf.


[^0]:    ${ }^{1}$ Related developments are the indirect addilog system (Houthakker, 1960) and the generalised addilog system (Bewley, 1982).

[^1]:    The last (most frequent) category in each variable acts as reference.

    * t-test p-value $<0.01$; ** t -test p -value $<0.001$

[^2]:    ${ }^{1}$ Minimum Covariance Determinant

[^3]:    ${ }^{2}$ GSI

[^4]:    ${ }^{1}$ The absolute difference between subsequent count values for the selected elements is proportionally less than $7 \%$ of the average total of counts for $\mathrm{Ti}, \mathrm{Fe}, \mathrm{Si}, \mathrm{Ca}, \mathrm{K}$ and $\mathrm{Sr} ; 14 \%$ for Al , and $24 \%$ for S .

[^5]:    ${ }^{2}$ Further trace elements were not considered herein because they have scattered information, more than $10 \%$ of values under the detection limit (rounded zeros), or the rounded zeros appear grouped, which require a different approach to replace the zeros.

