PROCEEDINGS OF THE $5TH$ INTERNATIONAL WORKSHOP ON COMPOSITIONAL DATA ANALYSIS

CoDaWork 2013
June 3-7, 2013, Vorau, Austria

FULL PAPERS

K. HRON, P. FILZMOSER AND M. TEMPL (EDS.)

Sponsors:

ISBN: 978-3-200-03103-6

Dear Friends and Colleagues,

We warmly welcome you to Vorau, for the Fifth International Workshop on Compositional Data Analysis (CoDaWork 2013). As many of you know, this workshop has been established as a forum of discussion for people concerned with the statistical treatment and modeling of compositional data or other constrained data sets, and the interpretation of models or applications involving them. The presented Workshop Proceedings consist of 20 papers of a wide range from theoretical issues to various applications of compositional data analysis.

In order to better understand different methodological aspects, how the compositional data are often analyzed, we need to know the complicated history of compositional data analysis. There, more than in many other mathematical fields, the development of a concise approach for a reasonable data treatment was the result of an extensive effort during the last century. The first comprehensive introduction to the analysis of compositional data using the logratio methodology is the monograph *Statistical Analysis of Compositional Data* (1986) by John Aitchison. Since then, much work has been devoted to the theoretical development of the field and to its applications in practice. The staying-in-the simplex approach that uses the specific algebraic-geometric structure of the simplex to express compositions using proper coordinates as standard real data, turned out to provide a solution to many statistical problems with an application potential in many fields where compositional data naturally arise. However, experience from the last three decades shows that not only a comprehensive theory, but rather fruitful applications justify the recent position of the logratio methodology as a reasonable alternative to the other existing approaches to compositional data analysis. Successful application areas are for example geochemistry, analytical chemistry, biological sciences, economics, but also and many others.

We shall continue to think about possible future research lines and relations to the other mathematical (statistical) fields. Obviously, a better understanding between the “traditional approach” to compositional data analysis (represented by a standard analysis of constrained data on the simplex) and the logratio methodology (that uses specific geometrical properties of data carrying relative information, with their possible representation as constrained data) could help to enrich both of them and prevent to form closed groups, competing with each other. A novel link between that relative and absolute information, carried by compositional data, is introduced using the product space $\mathcal{T}$, a tool for compositional data with a total. Contributions, related to regression and association issues, compositional (contingency) tables, high-dimensional compositions, calculus and the corresponding visualization techniques show a strong potential of compositional data analysis to handle problems that occur frequently in practice. Further, the presented biological, ecological, chemometrical and economical applications point out concrete cases, where the compositional methodology is particularly useful. Simultaneously, they turn us also back to roots and inspire to search primarily for answers, asked by practicians from both the traditional and new-coming fields, where compositional data form many of experimental data sets. Finally, all the proceedings papers aim to promote a special care that needs to be undertaken for a reasonable analysis of compositional data instead of ignoring the problem or hoping for a simple overall solution.

In order to further propagate leading ideas of CoDaWork 2013, authors of selected papers (the selection will be done by the CoDaWork Scientific Program Committee) will be asked after CoDaWork 2013 to submit an extended version of the paper to a special issue of the journal “Statistical Modelling”. The papers for this special issue will be peer-reviewed.

The proceedings papers reflect the highlights of CoDaWork 2013 and the compositional data analysis in general. Accordingly, the Scientific Committee has endeavored to provide a balanced and stimulating program that will appeal to the diverse interests of the participants. We keep to the tradition in the series of workshops on compositional data analysis that no parallel sessions are allowed in order to enable a knowledge transfer as well as to provoke a fruitful discussion among researchers from various fields of interest in the compositional data community. After the previous four CoDaWorks that were successfully organized by the “Girona gang” from the University of Girona in Spain, now for the first time the conference venue takes a new place. The Local Organizers from the “Vienna-Olomouc group” hopes that Vorau, a small village with rich history in the pictorial Styrian countryside, the “green heart of Austria”, will provide the appropriate environment to enhance your contacts and to establish new ones.

We acknowledge work of all organizers and the support of our hosts and sponsors, and particularly the Vienna University of Technology, the Austrian Statistical Society (ÖSG), the International Association for Mathematical Geology (IAMG), the United Nations Industrial Development Organization (UNIDO), and the private company data-analysis OG.

We wish you a productive, stimulating workshop and a memorable stay in Vorau.

Vorau, June 3, 2013

Karel Hron, Peter Filzmoser and Matthias Templ
(Editors)

Contents

The risk ratio versus odds ratio argument revisited from a compositional data analysis perspective

J. BACON-SHONE

Social Sciences Research Centre, The University of Hong Kong, Hong Kong, johnbs@hku.hk

1. Introduction

Most statisticians, geologists and economists working with compositional data would now accept that the argument about whether to use log-ratios when modelling compositional data has been won by those following the approach of Aitchison (1985). It is clear that the Euclidean metric is not an appropriate measure for data in the simplex and should be replaced by the Aitchison distance. However, there are many other scientific disciplines that have not heard the message and continue to use distance metrics that are inappropriate for the sample space. In particular, when the composition is not directly observable, but is a vector of parameters in a model, there continues to be confusion as to what is an appropriate model. For example, the argument amongst epidemiologists and statisticians about whether to use risk ratios or odds ratios has raged for more than 30 years (Cummings, 2009). Those favouring risk ratios highlight the ease of interpretation by clinicians and that the risk ratio is not affected when adjustment is made by a variable that is not a confounder. Those favouring odds ratios point out that odds ratios are symmetrical with respect to both the outcome and risk variables, which is consistent with the likelihood ratio principle, unlike risk ratios.

The specific situation that I wish to examine here is when the composition is a set of (unobservable) probabilities. We will start with the very simple situation of two complementary probabilities representing the chance of success and failure commonly used in many epidemiological models. In these models, the aim is to understand the effect of a range of factors on the chance of death or survival.

2. A simple example

Consider the simplest situation, where we have a single factor X that is present or absent and we observe how many individuals with or without X present are alive or dead.

If p_1 is the probability of being alive with X present (labelled as 1) and p_0 is the corresponding probability when X is not present (labelled as 0) and n_1 and n_0 are the corresponding numbers of individuals found alive, while m_1 and m_0 are the corresponding numbers found dead, then the log likelihood function is clearly:

$$LL=n_1\log(p_1) + m_1\log(1-p_1) + n_0\log(p_0) + m_0\log(1-p_0)$$

In this situation, there is no problem finding a point estimate for the risk ratio p_1/p_0, for which the maximum likelihood estimate is:

$$n_1(n_0+m_0)/(n_0(n_1+m_1)) ;$$

or for the odds ratio, which is $p_1(1-p_0)/(p_0(1-p_1))$, for which the maximum likelihood estimate is:

$n_1m_0/(m_1n_0)$.

Using likelihood ratio or Bayesian analysis, it is also straightforward to construct interval estimates for both measures.

Constructing exact confidence intervals for the risk ratio is difficult, although for the odds ratio it is straightforward after conditioning. As a result, it is common to use bootstrap intervals based on the maximum likelihood estimates.

3. The arguments

Cummings (1979) nicely summarizes most of the arguments between use of the risk ratio and odds ratio as follows:

3.1 Interpretation overall

The argument states that risk ratio is superior because it is more easily understood and used by clinicians. The supporting evidence is a long list of papers where the authors have clearly misunderstood the difference between risk ratio and odds ratio. In practice, if the risk is low under all scenarios, then there is little difference between the two measures. However, if the risk exceeds 10%, then the error in using the wrong measure is arguably significant.

3.2 Interpretability of averages

The estimate of the risk ratio (unlike odds ratio) will not change if we adjust for a variable that is not a confounder. This again means that it is easier to interpret a risk ratio, although if there are any confounders, this advantage disappears.

3.3 Constancy of odds ratios

As odds ratios are from R^+, it is possible for an odds ratio to be constant across a population, whereas because of the constraint on the risk ratio, this is impossible as there is an upper limit on probabilities, so the risk ratio cannot exceed the reciprocal of the unexposed risk. In short, risk ratio modelling does not address the implicit constraint. This argument should sound familiar to compositional data analysts and we will re-examine this argument in more detail below.

3.4 Symmetry of odds ratios

When calculating odds ratios, it makes no difference how we label outcome, we obtain the same result. However, risk ratios change if we change the labelling.

3.5 Estimation problems with risk ratios

Cummings (1979) does not discuss this problem, but Williamson (2011) devotes his entire thesis to discussing in detail how to address the problem that maximum likelihood methods often fail to converge for log binomial models, which involve fitting binomial models with linear models for the log risk ratio, which are the logical extension of the simple model considered above. He shows that for many methods commonly used for modelling risk ratios, there are problems with estimation using standard iterative methods because the likelihood maximum is on the boundary, as a consequence of the constraint on the probabilities.

4. More complex situations

For more complex scenarios, it is common to frame the problem in terms of generalized linear models, popularized in the book by Nelder and McCullagh (1989). In this formulation, the underlying distribution is that the outcome variables follow independent Bernoulli distributions with probability p_i of success, with a linear predictor

$\mu = X \beta$

where $\mu = g(p)$

for some link function g() that is a monotonic differentiable function.

The most common approach is to use the canonical link, which ensures that the linear predictor μ yields $X^T Y$ as the sufficient statistic, which in this case means using g(p)=log(p/(1-p)).

Other possibilities considered by Nelder and McCullagh for the case of binary data are that g(p) is

log(-log(1-p))

-log(-log(p))

or

$\Phi^{-1}(p)$.

The models being used by those modelling risk ratios directly, however, correspond to g(p) is

log(p).

5. Discussion

What then does compositional data analysis have to offer, in helping to identify a resolution for this longstanding argument?

First, it will be recalled that the early arguments against log ratio analysis included claims that it must be faulty because it was harder to interpret than linear model analysis (or than mappings onto the sphere rather than R^d). It was only when people understood that some questions do not make sense for compositional data, such as trying to model the difference in compositions using Euclidean distance that progress could be made.

Second, we now recognise the need to either map all our questions onto R^d so that we can use the Euclidean metric or alternatively define a distance metric that is appropriate for the original sample space.

Let us now re-examine the difference between log binomial regression (models for risk ratios) and logistic regression (models for odds ratios)

As noted above, the standard log binomial linear model is based on the link function:

$g(p) = \log(p)$

Whereas for logistic regression, the matching link function is:

$g(p)=\log(p/(1-p))$.

On the face of it, there may seem no theoretical reason why one model should be superior to the other.

However, p and 1-p comprise a simple composition, even if p is a parameter, rather than data. Hence, it makes sense to a compositional data analyst that we must either use logistic regression with a Euclidean metric or equivalently model p on the simplex using the Aitchison metric.

If we examine the log binomial model, there is an obvious mapping problem because log(p) only covers R^-, so there is an implicit constraint on $X\beta$ to cover only R^- as well. This constraint is very problematic because as a constraint on β, it depends on X. This means that if we do two experiments with different X, then the constraints on β are different, so there is no simple way to combine our results. This also means that regardless of sample size, adding one new observation with a different value of X can significantly change the constraint for β and hence the estimate.

It is interesting to note that Nelder and McCullagh mention that all their suggestions for g(p) (unlike g(p)=log(p)) correspond to "inverses of well-known cumulative distribution functions having support on the entire real axis", although they do not explain why this matters, presumably because they consider this so glaringly obvious. They also go on to explain that the logistic function has the key advantage over other link functions of giving the same answer for prospective and retrospective sampling (i.e. conditioning on either row or column totals).

In short, it is impossible to frame sensible questions about β for a log binomial model unless there are fixed boundaries for X. In other words, we not only must know the sample space for Y, but must also have a finite sample space for X and cannot generate sensible models if the sample space for X is continuous.

This is clearly a very serious weakness of log binomial models. The advantage of risk ratios being easier to interpret is far outweighed by the difficulty of interpreting the parameters of the underlying models other than for very simple situations. Constructing linear models with constraints on the parameter space that depend on X does not seem sensible.

6. Conclusions

By reviewing the arguments about risk ratio versus odds ratio from a compositional data analysis perspective, it is clear that the log binomial models that underpin risk ratio models have serious flaws and cannot make sense if the sample space for X is not finite. In comparison, the logistic regression models that underpin odds ratio models are consistent with compositional data analysis principles and can handle any sample space for X, even if the consequence is models that may appear harder to interpret for clinicians.

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.

Aitchison, J. and J. Bacon-Shone (1981). Bayesian risk ratio analysis. American Statistician: 254-257.

Cummings, P. (2009). The relative merits of risk ratios and odds ratios. Arch. Pediatr. Adoles. Med. 163 (5), 438-445.

McCullagh, P. and J.A. Nelder. (1989). Generalized Linear Models (second edition), Chapman and Hall, London.

Williamson, T.S. (2011). Log-Binomial Models: Maximum Likelihood and Failed Convergence, PhD thesis, University of Calgary, Calgary.

Compositional regression with unobserved components or below detection limit values

K.G. VAN DEN BOOGAART[1,2], R. TOLOSANA-DELGADO[1], K. HRON[3], M. TEMPL[4] and P. FILZMOSER[4]

[1]Department of Modelling and Valuation - Helmholtz Institute Freiberg for Resources Technology, Germany
[2]Institute for Stochastics - Technical University Bergakademie Freiberg, Germany
[3]Department of Mathematical Analysis and Applications of Mathematics, and Department of Geoinformatics - Palacký University Olomouc, Czeck Republich
[4]Department of Statistics and Probability Theory - Vienna University of Technology, Vienna, Austria
boogaart@hzdr.de

Abstract

The typical way to deal with zeroes and missing values in compositional data sets is to impute them with a reasonable value, and then the desired statistical model is estimated with the imputed data set, e.g. a regression model. This contribution aims at presenting alternative approaches to this problem within the framework of Bayesian regression with a compositional response. In a first step, a compositional data set with missing data is considered to follow a normal distribution on the simplex, which mean value is given as an Aitchison affine linear combination of some fully-observed explanatory variables. Both the coefficients of this linear combination and the missing values can be estimated with standard Gibbs sampling techniques. In a second step, a normally-distributed additive error is considered superimposed on the compositional response, and values are taken as "below the detection limit" (BDLs) if they are "too small" in comparison with the additive standard deviation of each variable (usually, a 3σ rule is applied here). Within this framework, the regression parameters and all missing values (including BDLs) can be estimated, albeit this time with a less efficient Metropolis-Hastings algorithm. Both methods estimate the regression coefficients without need of any preliminary imputation step, and adequately propagate the uncertainty derived from the fact that the missing values and BDLs are not actually observed, something imputation methods cannot achieve.

1 Introduction

Rounded zeroes and missing values in compositional data are often treated as "sick" observations, that must be "cured" (say, replaced) before one can apply the log-ratio methodology. This replacement or *imputation* strategy is often reasonable, when the number of such irregular data is small and if the analyst is seeking a "quick and dirty" answer. However, whenever the number of irregularities increases, or one is interested in closely monitoring the uncertainty of the fitted models, then the imputation may severely undererstimate the uncertainty. This problem is specially critical, for instance, when testing the significance of some regression parameters.

The regression of a compositional variable against a set of covariables is an important tool to analyse the dependence of compositions to external influences. It can essentially be treated with the principle of working in coordinates (*take log-ratios, analyse the scores, back-transform the coefficients*), by using a multivariate regression model for the additive or isometric log ratio transform, using available standard software (e.g. R). Actually, this can be seen as a maximum likelihood (ML) estimation problem, where the compositional response is assumed to follow an additive logistic normal distribution (ALN), which conditional expectation is given by a linear combination of the explanatory variables.

But when the compositional response data has several missing values, the log-ratio approach fails, because missing components do not simply correspond to missing log-ratios. This renders compositional regression unavailable for many existing datasets, like e.g. containing incomplete analysis data (classically known as *missing values* or very low concentrations in individual components (*rounded zeroes* or *values below the detection limit*).

Within this framework, this contribution proposes an integrated approach, where each incomplete observation contributes with an incompletely determined likelihood to the estimation process. For missing values, this implies that the likelihood of that observation is defined in a subspace only (the subspace associated to the observed subcomposition).

Observations below detection limit (BDL) are more complicated. If one assumes that it is still an outcome of an ALN, a BDL has a likelihood that can be computed quasi-analytically, from

the normal cumulative distribution function (Palarea-Albaladejo et al., 2007). However, for most chemical data, the meaning of a BDL is a bit more complex, involving some additive error on top of the ALN distribution (van den Boogaart et al., 2011). This problem can also be solved, making use of the concept of latent variables. Beyond modelling BDLs and missing values, this model is able to consider additive errors even to the observed subcomposition, and can thus be of more general use in case of polluted data sets. On the downside, it requires extensive Markov Chain Monte Carlo computing in a Bayesian approach.

Combining both strategies as needed, we can fit compositional regression models efficiently in case of missing data. The approach also provides a complete "imputed" dataset, distributed according to the true conditional distribution (given all the data) of the unobserved true compositions, filtering the measurement error out.

Section 2 introduces the geometric and statistical concepts necesssary to work with compositional data. Section 3 presents two adaptations of classical Bayesian multivariate regression to deal with a compositional response, with and without additive error. Section 4 briefly reviews the kinds of missing values, and gives the necessary modifications to the regression models to work with them. Some conclusions close this contribution in section 5.

2 Notation and grounding concepts

A D-component vector $\mathbf{z} = [z_1, z_2, \ldots, z_D]$ is considered a composition if its components show the relative importance of D parts forming a total. The key property of a compositional vector is its scale invariance (Aitchison, 1986): because the only relevant information of a composition is relative, scaling it by any constant value does not change its meaning, i.e. $p \cdot \mathbf{z} \equiv \mathbf{z}$. The sample space of compositional data is the simplex, $\mathcal{S}^D$, defined as the positive orthant of the D-dimensional real space $\mathbb{R}^D_+$ equipped with the scale invariance eqivalence (Barceló-Vidal et al., 2001). For practical reasons, however, this definition is simplified to the sample space of representatives, i.e. the set of non-negative components and total sum equal to 1,

$$\mathcal{S}^D = \left\{ \mathbf{z} \,\middle|\, z_i \geq 0 \wedge \sum_{i=1}^{D} z_i = 1 \right\}.$$

If a composition does not satisfy the constant sum constraint, it can be forced to do so by the application of the closure operation

$$\mathbf{z}' = \mathcal{C}[\mathbf{z}] = \frac{1}{\sum_{i=1}^{D} z_i} \cdot \mathbf{z},$$

without loss of any relevant information. Pawlowsky-Glahn and Egozcue (2001) showed that the simplex can be given a Euclidean vector space structure by the operations of: *perturbation* (closed component-wise product of two compositions), *powering* (closed component-wise powering of a composition by a scalar) and the *Aitchison scalar product* (proportional to the scalar product of the vectors formed by all possible pairwise logratios of the two compositions).

A *subcomposition* is a composition built from another composition by selecting a subset of its parts, and considering it scale invariant, i.e. applying the closure operation to the subset of chosen components. Within the Euclidean geometry of the simplex, the sample space of a subcomposition is a vector subspace of the larger D-part simplex.

The scale invariance condition implies that: (1) the total sum of all components is in general an irrelevant quantity, and (2) all relevant information is conveyed by the component ratios, e.g., z_i/z_j. The Euclidean structure of the simplex allows to identify any composition with its vector of $(D-1)$ coordinates in any arbitrary reference basis. Actually, these coordinates are always balanced logratios or log-contrasts, linear combinations of log-transformed components which coefficients add up to zero,

$$\zeta = \mathbf{v}^t \cdot \ln \mathbf{z}, \qquad \mathbf{v}^t \cdot \mathbf{1}_D = 0,$$

with the logarithm applied component-wise. The vector $\mathbf{1}_D$ denotes a vector of D ones. If $(D-1)$ linearly independent vectors are considered, then a basis is obtained,

$$\boldsymbol{\zeta} = \mathbf{V}^t \cdot \ln \mathbf{z}, \qquad \mathbf{V}^t \cdot \mathbf{1}_D = \mathbf{0}_{D-1}.$$

For practical reasons, one can also assume an orthogonal reference basis, i.e. where $\mathbf{V}^t \cdot \mathbf{V} = \mathbf{I}_{D-1}$ the $(D-1)$-identity matrix, where superindex t denotes transposition. In this case, one defines the so-called isometric logratio transformation, linking compositions and coordinates through

$$\mathrm{ilr}(\mathbf{z}) := \mathbf{V}^t \cdot \ln \mathbf{z} = \boldsymbol{\zeta}; \quad \mathrm{ilr}^{-1}(\boldsymbol{\zeta}) := \mathcal{C}[\exp(\mathbf{V} \cdot \boldsymbol{\zeta})] = \mathbf{z}. \tag{1}$$

Non-orthogonal bases can also be used, but then the inverse transformation expression involves a generalized inversion of $\mathbf{V}^t$, and is not so straightforward.

An alternative log-ratio representation is provided by the *centered logratio transformation* (clr, Aitchison, 1986), a one-to-one transformation defined as

$$\mathrm{clr}(\mathbf{z}) := \ln \frac{\mathbf{z}}{\sqrt[D]{z_1 z_2 \cdots z_D}} = \boldsymbol{\zeta}_* \quad \mathrm{clr}^{-1}(\boldsymbol{\zeta}_*) := \mathcal{C}[\exp(\boldsymbol{\zeta}_*)] = \mathbf{z}.$$

Between clr and ilr transformations there is also a one-to-one equivalence,

$$\mathrm{ilr}(\mathbf{z}) = \mathbf{V}^t \cdot \mathrm{clr}(\mathbf{z}), \quad \mathrm{clr}(\mathbf{z}) = \mathbf{V} \cdot \mathrm{ilr}(\mathbf{z}), \tag{2}$$

which only holds for orthonormal bases.

A set of coordinates can be built to capture a certain subcomposition. Assuming for simplicity that the subcomposition contains the first S of the D components, then the vector of coordinates should have:

1. first, $S-1$ logcontrasts between the S components of the subcomposition,
2. second, $D-S-1$ logcontrasts between the remaining $D-S$ components,
3. finally, a balance between the geometric means of the components of the two subcompositions,

$$\zeta_{D-1} = \sqrt{\frac{S(D-S)}{D}} \ln \frac{\sqrt[S]{\prod_{i=1}^{S} z_i}}{\sqrt[D-S]{\prod_{i=S+1}^{D} z_i}}.$$

For the goals of this contribution, the sets of logcontrasts within the two subcompositions are irrelevant. Egozcue and Pawlowsky-Glahn (2005) give expressions and rules to build these subbases. We just denote by $\mathbf{V}_o$ and $\mathbf{V}_m$ the columns of the matrix $\mathbf{V}$ respectively linked to the observed subcomposition and to the non-observed subcomposition and balance. Thus, $\mathbf{V}_o$ has $D-S-1$ columns, and $\mathbf{V}_m$ has S columns.

A D component random composition will be denoted as $\mathbf{Z}$. It will be said to have an *additive logistic normal distribution* (ALN Aitchison, 1982), also known as *normal distribution on the simplex* (Mateu-Figueras et al., 2003), if any of its vector of log-ratio coordinates has a joint $(D-1)$-variate normal distribution. The normal distribution on the simplex has as density function

$$f_{\mathbf{Z}}(\mathbf{z}|\boldsymbol{\mu}_{\mathbf{V}}, \boldsymbol{\Sigma}_{\mathbf{V}}) = \frac{1}{\sqrt{(2\pi)^{D-1}|\boldsymbol{\Sigma}_{\mathbf{V}}|}} \cdot \exp\left[-\frac{1}{2}\left(\mathbf{V}^t \cdot \ln \mathbf{z} - \boldsymbol{\mu}_{\mathbf{V}}\right)^t \cdot \boldsymbol{\Sigma}_{\mathbf{V}}^{-1} \cdot \left(\mathbf{V}^t \cdot \ln \mathbf{z} - \boldsymbol{\mu}_{\mathbf{V}}\right)\right] \tag{3}$$

As happens with the lognormal distribution, the parameters of the ALN distribution will be identified with those of the normal distribution of the transformed scores, i.e. the mean vector $\boldsymbol{\mu}_{\mathbf{V}}$ and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{V}}$ of the ilr-transformed composition, using matrix $\mathbf{V}$ in Eq. (1). Alternatively, we can also use an expression based on the clr transformed scores, thanks to Eq. (2), and an equivalent expression for the covariance $\boldsymbol{\Sigma}$ of clr-transformed data

$$\boldsymbol{\Sigma} = \mathbf{V}^- \cdot \boldsymbol{\Sigma}_{\mathbf{V}} \cdot \mathbf{V}^{-t}, \quad \boldsymbol{\Sigma}_{\mathbf{V}} = \mathbf{V}^t \cdot \boldsymbol{\Sigma} \cdot \mathbf{V}, \tag{4}$$

where $\mathbf{V}^-$ is the Moore-Penrose generalized inverse of $\mathbf{V}$, and $\mathbf{V}^{-t}$ the transposed generalized inverse. Inverting $\boldsymbol{\Sigma}$ is tricky, because this matrix is always singular and therefore has no inverse and its determinant is zero, strictly speaking. Both problems are bypassed by considering again Moore-Penrose generalized inverse, which in this case gives

$$\boldsymbol{\Sigma}^- = \mathbf{V}^- \cdot \boldsymbol{\Sigma}_{\mathbf{V}}^{-1} \cdot \mathbf{V}^{-t}, \tag{5}$$

which happens to be the same whichever basis is used for the calculations. Note that both matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^-$ sum up to zero by rows and by columns. Note as well that, if $\mathbf{V}$ is an orthogonal matrix, then $\mathbf{V}^- = \mathbf{V}^t$ and $\mathbf{V}^{-t} = \mathbf{V}$, which simplifies a bit the expressions. For the sake of simplicity, we assume such an orthogonal matrix from this point on.

In the same way as an inverse can be generalized, the determinant of $\boldsymbol{\Sigma}$ can be defined as the determinant of any of its ilr representations through Eq. (4), i.e. $|\boldsymbol{\Sigma}| := |\boldsymbol{\Sigma}_{\mathbf{V}}|$, as this is an invariant function. Moreover, $|\boldsymbol{\Sigma}| = 1/|\boldsymbol{\Sigma}_{\mathbf{V}}|^{-1} = 1/|\boldsymbol{\Sigma}|^-$. With these definitions

$$f_{\mathbf{Z}}(\mathbf{z}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sqrt{\frac{|\boldsymbol{\Sigma}^-|}{(2\pi)^{D-1}}} \cdot \exp\left[-\frac{1}{2}\left(\mathrm{clr}(\mathbf{z}) - \boldsymbol{\mu}\right)^t \cdot \boldsymbol{\Sigma}^- \cdot \left(\mathrm{clr}(\mathbf{z}) - \boldsymbol{\mu}\right)\right]. \tag{6}$$

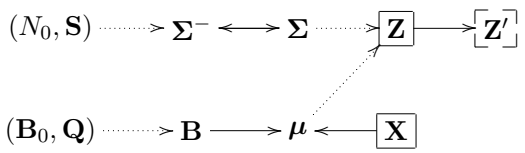


Figure 1: Scheme of relation between parameters, variables and observables (within a framebox) for a compositional Bayesian regression model. Dotted lines represent random relations, solid lines deterministic ones. The observable $\mathbf{Z}'$ will appear in the case of missing values, and in that case $\mathbf{Z}$ will be considered a latent, unobservable variable.

3 Bayesian regression with compositional response

3.1 Compositional Bayesian regression

In a conventional multivariate regression model, a response is considered a function of some explanatory variables $\mathbf{X}$ (usually, including a constant predictor 1 that will later account for the intercept), in such a way that a linear combination of the explanatory variables gives the expected value of the response. This response is then considered joint normally distributed with an unknown constant variance. This can be adapted straight to compositional response, as

$$\operatorname{ilr}(\mathbf{Z})|\mathbf{X}, \boldsymbol{\Sigma}_\mathbf{V} \sim \mathcal{N}(\mathbf{B}_\mathbf{V} \cdot \mathbf{X}, \boldsymbol{\Sigma}_\mathbf{V}), \tag{7}$$

which has the slight inconvenience to depend on the ilr basis chosen. It is more practical to describe the regression parameters in relation to the clr transformation,

$$\mathbf{B}_\mathbf{V} \cdot \mathbf{X} = \mathrm{E}\left[\operatorname{ilr}(\mathbf{Z})\right] = \mathrm{E}\left[\mathbf{V}^t \cdot \ln \mathbf{z}\right] = \mathbf{V}^t \cdot \mathrm{E}\left[\ln \mathbf{z}\right] = \mathbf{V}^t \cdot \mathrm{E}\left[\operatorname{clr}(\mathbf{z})\right],$$

which gives a basis independent representation of the coefficients as $\mathbf{B} = \mathbf{V} \cdot \mathbf{B}_\mathbf{V}$, with inverse relation $\mathbf{B}_\mathbf{V} = \mathbf{V}^t \cdot \mathbf{B}$. These expressions do not depend on the basis used, though that inverse expression requires the basis to be orthogonal. Note that the columns of $\mathbf{B}$ must sum up to zero.

Given an observed pair $(\mathbf{x}_n, \mathbf{z}_n)$, the likelihood of the regression coefficients and residual variance is then

$$L(\mathbf{B}, \boldsymbol{\Sigma}^-|\mathbf{x}_n, \mathbf{z}_n) \propto f_\mathbf{Z}(\mathbf{z}_n|\mathbf{B} \cdot \mathbf{x}_n, \boldsymbol{\Sigma}),$$

or taking logs and using expression (6)

$$\begin{aligned} l(\mathbf{B}, \boldsymbol{\Sigma}^-|\mathbf{x}_n, \mathbf{z}_n) &= \kappa - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\operatorname{ilr}(\mathbf{z}_n) - \mathbf{V}^t \cdot \mathbf{B} \cdot \mathbf{x}_n)^t \cdot \boldsymbol{\Sigma}_\mathbf{V}^{-1} \cdot (\operatorname{ilr}(\mathbf{z}_n) - \mathbf{V}^t \cdot \mathbf{B} \cdot \mathbf{x}_n) \quad (8) \\ &= \kappa + \frac{1}{2}\ln|\boldsymbol{\Sigma}^-| - \frac{1}{2}(\operatorname{clr}(\mathbf{z}_n) - \mathbf{B} \cdot \mathbf{x}_n)^t \cdot \boldsymbol{\Sigma}^- \cdot (\operatorname{clr}(\mathbf{z}_n) - \mathbf{B} \cdot \mathbf{x}_n). \quad (9) \end{aligned}$$

If a sample of pairs is available, $(\mathbf{x}_n, \mathbf{z}_n); n = 1, 2, \ldots, N$, then the joint log-likelihood of the sample is

$$\begin{aligned} l(\mathbf{B}, \boldsymbol{\Sigma}^-|\mathbf{x}_n, \mathbf{z}_n; n &= 1, 2, \ldots, N) = \sum_{n=1}^{N} l(\mathbf{B}, \boldsymbol{\Sigma}^-|\mathbf{x}_n, \mathbf{z}_n) = \quad (10) \\ &= \kappa' + \frac{N}{2}\ln|\boldsymbol{\Sigma}^-| - \frac{1}{2}\sum_{n=1}^{N}(\operatorname{clr}(\mathbf{z}_n) - \mathbf{B} \cdot \mathbf{x}_n)^t \cdot \boldsymbol{\Sigma}^- \cdot (\operatorname{clr}(\mathbf{z}_n) - \mathbf{B} \cdot \mathbf{x}_n). \end{aligned}$$

If the pairs of explanatory-explained are ordered column-wise in matrices $\mathbf{X}$ and $\mathbf{Z}_*$ (this last where each column is a clr transformed composition), then maximum likelihood estimates of the parameters are provided by

$$\hat{\mathbf{B}} = (\mathbf{Z}_* \cdot \mathbf{X}^t) \cdot (\mathbf{X} \cdot \mathbf{X}^t)^{-1}, \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{N}(\mathbf{Z}_* - \hat{\mathbf{B}} \cdot \mathbf{X}) \cdot (\mathbf{Z}_* - \hat{\mathbf{B}} \cdot \mathbf{X})^t.$$

A classical Bayesian approach to regression assumes certain prior, independent distributions for the parameters $\mathbf{B}$ and $\boldsymbol{\Sigma}$, which are combined with the likelihood (Eq. 10) through Bayes Theorem to give the joint posterior distribution of the parameters,

$$\pi[\mathbf{B}, \boldsymbol{\Sigma}|\mathbf{X}, \mathbf{Z}] \propto \pi^0[\mathbf{B}] \cdot \pi^0[\boldsymbol{\Sigma}] \cdot \exp(l(\mathbf{B}, \boldsymbol{\Sigma}|\mathbf{x}_n, \mathbf{z}_n; n = 1, 2, \ldots, N)).$$

Because of practical reasons, it is common to choose these prior distributions $\pi^0[\cdot]$ as the conjugate priors of the normal distribution:

- $\mathbf{B}$ is assumed to follow a (degenerate) normal distribution with mean value $\mathbf{B}_0$ and precision matrix $\mathbf{Q}$; equivalently, that means that $\mathbf{B}_\mathbf{V}^t$ follows a normal distribution with mean value $\mathbf{V}^t \cdot \mathbf{B}_0$ and precision matrix $\mathbf{Q}_\mathbf{V} = (\mathbf{1}_P \otimes \mathbf{V}^t) \cdot \mathbf{Q} \cdot (\mathbf{1}_P \otimes \mathbf{V}^t)^t$ (with the matrix $\mathbf{B}$ converted to a vector by stacking it column-wise), with P the number of predictor variables (eventually including the constant). This has an associated covariance matrix of $\mathbf{Q}_\mathbf{V}^{-1}/2$.
- $\mathbf{\Sigma}^-$ is assumed to follow a Wishart distribution, with size N_0 and variance parameter $\mathbf{S}$; this can also be expressed in any particular ilr base as $\mathbf{S}_\mathbf{V} = (\mathbf{1}_P \otimes \mathbf{V}^t) \cdot \mathbf{S} \cdot (\mathbf{1}_P \otimes \mathbf{V}^t)^t$.

These prior specifications give rise to the following posterior *conditional* distributions,

$$\mathbf{\Sigma}^-|\mathbf{X},\mathbf{Z},\mathbf{B} \sim \mathcal{W}\left(N+N_0, (\mathbf{S}^- + \hat{\mathbf{\Sigma}}^-)^-\right), \tag{11}$$

$$\mathbf{B}|\mathbf{X},\mathbf{Z},\mathbf{\Sigma}^- \sim \mathcal{N}_{P(D-1)}\left((\mathbf{Q}+\mathbf{R})^{-1} \cdot (\mathbf{Q} \cdot c(\mathbf{B}_0) + \mathbf{R} \cdot c(\hat{\mathbf{B}})); \mathbf{Q}+\mathbf{R}\right), \tag{12}$$

with $\mathbf{R} = (\mathbf{X} \cdot \mathbf{X}^t) \otimes \hat{\mathbf{\Sigma}}^-$, denoting by $c(\mathbf{A})$ the column-wise stacking of matrix $\mathbf{A}$ in a vector, and specifying the normal distribution in terms of its precision matrix instead of its covariance.

Given that the posterior distribution is specified with its conditional distributions, it is suitable to explore it with a Gibbs sampling scheme. With the following algorithm, a sample of the posterior distribution will be obtained:

1. Fix the matrix of regression coefficients to a suitable value $\mathbf{B}^{(0)}$, e.g. intercept equal to a reasonable mean of the compositional response, slopes all equal to 0;
2. simulate a random value $(\mathbf{\Sigma}^-)^{(k+1)}$ out of the distribution (Eq. 11) of $\mathbf{\Sigma}^-|\mathbf{X},\mathbf{Z},\mathbf{B}^{(k)}$;
3. simulate a random value $\mathbf{B}^{(k+1)}$ out of the distribution (Eq. 12) of $\mathbf{B}|\mathbf{X},\mathbf{Z},(\mathbf{\Sigma}^-)^{(k+1)}$;
4. return to step 2 until the number of simulations is large enough.

This Gibbs sampler reaches a stable distribution quite fast, thus a small burn-in period is enough.

3.2 Considering additive error

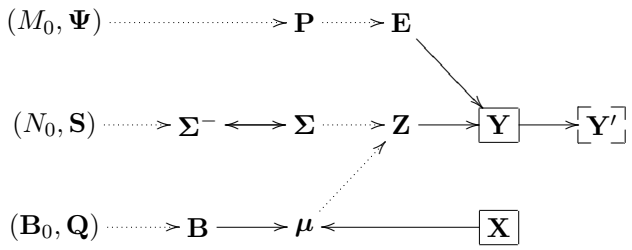


Figure 2: Scheme of relation between parameters, variables and observables (within a framebox) for a compositional Bayesian regression model with additive error. Dotted lines represent random relations, solid lines deterministic ones.

In chemical analysis, it is common to observe a composition with both a relative and an additive error, respectively derived from the common practice to compare samples with standards and the need to substract a noise background. In the preceding model, the relative error can be captured by $\mathbf{\Sigma}$. The additive error $\mathbf{E}$ is added in the model of Figure 2. Here, the composition itself is considered a latent parameter, which conditions the observations. These give

$$\mathbf{E}|\mathbf{P} \sim \mathcal{N}_D(\mathbf{0},\mathbf{P}), \qquad \mathbf{Y}|\mathbf{Z},\mathbf{P} \sim \mathcal{N}_D(\mathbf{Z},\mathbf{P}). \tag{13}$$

Note that $\mathbf{Y}$ is not anymore a composition in this model. Some information must be given with regard to $\mathbf{P}$: a full Bayesian framework would require a prior, which could be again considered an inverse Wishart distribution with size M_0 and variance $\mathbf{\Psi}$, as specified in fig. 2. However, we will consider it fixed, related to the analytical precision of the measuring machine. No analytical expression exists for the conditional distribution of $\mathbf{Z}|\mathbf{X},\mathbf{Y},\mathbf{B},\mathbf{\Sigma}^-,\mathbf{P}$. However, thanks to the conditional probability property $[A|B,C] \propto [A|B][C|A,B]$, where $[\cdot] = \pi[\cdot]$ denotes the probability distribution, we obtain

$$[\mathbf{Z}|\mathbf{X},\mathbf{Y},\mathbf{B},\mathbf{\Sigma}^-,\mathbf{P}] \propto [\mathbf{Z}|\mathbf{X},\mathbf{B},\mathbf{\Sigma}^-,\mathbf{P}] \cdot [\mathbf{Y}|\mathbf{X},\mathbf{Z},\mathbf{B},\mathbf{\Sigma}^-,\mathbf{P}] = [\mathbf{Z}|\mathbf{X},\mathbf{B},\mathbf{\Sigma}^-] \cdot [\mathbf{Y}|\mathbf{Z},\mathbf{P}]$$

The second step requires also the conditional independence properties implied in the scheme of figure 2, i.e. that given $\mathbf{Z}$, the blocks $(\mathbf{X}, \mathbf{B}, \mathbf{\Sigma}^-)$ and $(\mathbf{Y}, \mathbf{P})$ are conditionally independent. This is not closed, but it might still be suitable for a Metropolis-Hastings sampling scheme, as $[\mathbf{Z}|\mathbf{X}, \mathbf{B}, \mathbf{\Sigma}^-]$ follows an normal distribution on the simplex (Eq. 7), and $[\mathbf{Y}|\mathbf{Z}, \mathbf{P}] = [\mathbf{E}|\mathbf{P}]$ is a Normal distribution, according to Eq. (13). Thus, the following hybrid Gibbs/Metropolis-Hastings sampling scheme can be implemented:

1. Fix the matrix of regression coefficients to a suitable value $\mathbf{B}^{(0)}$, e.g. intercept equal to a reasonable mean of the compositional response, slopes all equal to 0; fix suitable latent compositions $\mathbf{Z}^{(0)}$, e.g. as the data $\mathbf{Y}$;
2. simulate a random value $(\mathbf{\Sigma}^-)^{(k+1)}$ out of the distribution $\mathbf{\Sigma}^-|\mathbf{X}, \mathbf{Z}^{(k)}, \mathbf{B}^{(k)}$ of Eq. (11);
3. simulate a random value $\mathbf{B}^{(k+1)}$ out of the distribution $\mathbf{B}|\mathbf{X}, \mathbf{Z}^{(k)}, (\mathbf{\Sigma}^-)^{(k+1)}$ of Eq. (12);
4. simulate $\mathbf{Z}^{(k+1)}$ with a Metropolis-Hastings algorithm;
   (a) simulate a candidate $\mathbf{Z}^*$ out of the distribution $[\mathbf{Z}|\mathbf{Y}, \mathbf{P}]$,
   (b) using Eq. (7), compute the transition probability
   $$p = \frac{[\mathbf{Z}^*|\mathbf{X}, \mathbf{B}^{(k+1)}, (\mathbf{\Sigma}^-)^{(k+1)}]}{[\mathbf{Z}^{(k)}|\mathbf{X}, \mathbf{B}^{(k+1)}, (\mathbf{\Sigma}^-)^{(k+1)}]},$$
   (c) take $\mathbf{Z}^{(k+1)} = \mathbf{Z}^*$ with probability $\min(1, p)$, otherwise take $\mathbf{Z}^{(k+1)} = \mathbf{Z}^{(k)}$;
   note that kind of scheme is an independent chain, according to Tierney (1994), which implies that the typical thinning necessary in the Metropolis-Hastings algorithm to ensure independent samples can be reduced to a minimum;
5. return to step 2 until the number of simulations is large enough.

In the Metropolis-Hastings step, we propose to simulate from $[\mathbf{Z}|\mathbf{Y}, \mathbf{P}]$ and accept/reject from $[\mathbf{Z}|\mathbf{X}, \mathbf{B}, (\mathbf{\Sigma}^-)]$ because the former will typically be much narrower than the latter.

4 Dealing with missing values

4.1 Values missing at random

To cope with the information derived from a missing value, it is important to know why the value was lost. A value can be lost due to a certain defined process, or just randomly. Roughly, the following classes of missingness are often considered

missing not at random (MNAR) the value is lost with a probability that might depend on the value itself and on other variables; this can only be solved with a model of the probability that each datum is missed/observed; values below the detection limit are an example of this;

missing at random (MAR) the value is lost with a probability that might only depend on other variables; if this probability can be modelled, then one could consider estimating this missing probability model at the same time that the main regression problem is tackled;

missing completely at random (MCAR) the value is lost with a probability independent of every variable.

Modelling general MNAR and MAR values is beyond the scope of this contribution, as they require to specify the mechanism of missingness.

MCAR values can be easily considered. Assume that the n-th datum has S missing variables. Then, the likelihood $l(\mathbf{B}, \mathbf{\Sigma}^-|\mathbf{x}_n, \mathbf{z}_n)$ of that datum can only be specified up to a subspace of dimension $(D-S-1)$. In particular, following the steps of definition of an ad-hoc ilr basis (Eq. 1) of section 2, we can choose an ilr that allows us to compute these $(D-S-1)$ coordinates, corresponding to the observed subcomposition. The rest (those within the S-component missing subcomposition and the balance) are missing values, which do not contribute to modify the likelihood in Eq. (8),

$$-2l(\mathbf{B}, \mathbf{\Sigma}^-|\mathbf{x}_n, \mathbf{z}_n) + \kappa = \ln|\mathbf{\Sigma}_{\mathbf{V}_o}| + (\ln(\mathbf{z}_n) - \mathbf{B}\cdot\mathbf{x}_n)^t \cdot \mathbf{V}_o \cdot \mathbf{\Sigma}_{\mathbf{V}_o}^{-1} \cdot \mathbf{V}_o^t \cdot (\ln(\mathbf{z}_n) - \mathbf{B}\cdot\mathbf{x}_n)$$

where the basis matrix $\mathbf{V}_o$ depends on which components of observation $\mathbf{z}_n$ are missing. Note as well that the determinant $|\mathbf{\Sigma}_{\mathbf{V}_o}|$ is computed as the product of the $D-S-1$ eigenvalues of $\mathbf{\Sigma}_{\mathbf{V}_o} = \mathbf{V}_o^t \cdot \mathbf{\Sigma} \cdot \mathbf{V}_o$.

If the model used is the one with additive error, then the problem of dealing with missing values is actually moved from $\mathbf{Z}$ to $\mathbf{Y}$, but the problem remains of the same type: the contribution of that observation to the likelihood is reduced.

In whichever layer that occurs, the result is that the conjugacy property is strictly lost (Gross and Torres-Quevedo, 1995). However, the latent variable approach allows to still sample the posterior distribution of the whole set of parameters and latent variables, by means of the same Markov Chain Monte Carlo schemes mentioned before (Gross, 2000).

4.2 Values below the detection limit

In general, a *value below the detection limit* or a *rounded zero* is a small value which cannot be distinguished from zero, given the accuracy of the measurement. If the measuring method is considered error-free, this concept can be easily accomodated into a Bayesian estimation of the scheme in figure 1, by considering that actually the composition is a partly observed, partly latent variable. A Bayesian estimation of this model has already been implemented via a Markov Chain Monte Carlo method (Palarea-Albaladejo et al., 2007), albeit based on an Expectation-Maximization algorithm applied to a non-orthogonal family of compositional coordinates, the additive logratio transformation (Aitchison, 1986). The key idea here is to consider a latent variable for each value, which will be replaced through the Gibbs chain by a value below the detection limit.

If the measuring method is considered error-prone (after figure 2), that model is not adequate. This is the most common case in chemical compositions, where chemical components are often determined by a comparing the readings obtained analysing the sample (signal, s) with those of a blank (background, b) and of a reference (r). Rougly, if the true concentration of the reference is known to be c_r, then the estimated concentration of the sample is $c_s = c_r(s-b)/(r-b)$. This structure induces a multiplicative error in the sample measurements (because of the scaling by c_r), but also a an additive error (because of the background subtraction). In this context, a value is reported below the detection limit if it cannot be statistically distinguished from the background (van den Boogaart et al., 2011). Considering this kind of definition, the key idea is to replace *in the data* the values below the detection limit *by the detection limit itself*, and let the model with additive error run.

4.3 Modified algorithms to deal with missing values

Both types of missing values are considered in the same way for a model without additive error (fig. 1). In this context, $\mathbf{Z}$ is considered a latent variable, and the observation is placed as an additional layer $\mathbf{Z}'$. Latent variables and observations are considered equal whenever the composition is fully observed, $\mathbf{z}_n = \mathbf{z}'_n$. But if a datum has any kind of missing values, then the $D-S$ observed components are considered the same for both vectors, $z_{in} = z'_{in}, i = S+1,\dots,D$, but different for the non-observed set (note that we assume here that the missing components were the first S). The simulation chain is modified then as follows:

1. Fix the matrix of regression coefficients to a suitable value $\mathbf{B}^{(0)}$, e.g. intercept equal to a reasonable mean of the compositional response, slopes all equal to 0; fix the latent variable $\mathbf{Z}^{(0)}$ to reasonable values, e.g. replace values below the detection limit by 2/3 of it, and missings at random by the compositional mean;
2. simulate a random value $(\mathbf{\Sigma}^-)^{(k+1)}$ out of the distribution (Eq. 11) of $\mathbf{\Sigma}^-|\mathbf{X},\mathbf{Z}^{(k)},\mathbf{B}^{(k)}$;
3. simulate a random value $\mathbf{B}^{(k+1)}$ out of the distribution (Eq. 12) of $\mathbf{B}|\mathbf{X},\mathbf{Z}^{(k)},(\mathbf{\Sigma}^-)^{(k+1)}$;
4. for each datum with missing values, simulate a random value $\mathbf{z}_n^{(k+1)}$ out of the distribution of $\mathbf{z}_n|\mathbf{x}_n,\mathbf{z}'_n,\mathbf{B}^{(k+1)},(\mathbf{\Sigma}^-)^{(k+1)}$; Palarea-Albaladejo et al. (2007) showed that this distribution is a normal distribution on the simplex,

$$\mathbf{z}_n \sim \mathcal{ALN}\left(\mathbf{B}^{(k+1)}\mathbf{x}_n, (\mathbf{\Sigma}^-)^{(k+1)}\right),$$

   which must be conditioned to the fact that $\mathbf{z}_n$ is partly observed by $\mathbf{z}'_n$. If an ilr basis is chosen following the steps of section 2 to separate missing from observed coordinates, then the distribution of the MAR is

$$\mathbf{V}_m \ln \mathbf{z}_n \sim \mathcal{ALN}\left((\mathbf{V}_m + (\mathbf{\Sigma}_{mo}\cdot\mathbf{\Sigma}_{oo}^{-1})^{(k+1)}\mathbf{V}_o)\mathbf{B}^{(k+1)}\mathbf{x}_n, (\mathbf{\Sigma}_{mm}-\mathbf{\Sigma}_{mo}\cdot\mathbf{\Sigma}_{oo}^{-1}\mathbf{\Sigma}_{om})^{(k+1)}\right).$$

   The distribution of the BDLs is the same, but further conditioned to correspond to a value below the detection limit (Palarea-Albaladejo et al., 2007).
5. Return to step 2 until the number of simulations is large enough.

In the model with additive error, exactly the same strategies apply to deal with the difference between the observed readings $\mathbf{Y}'$ and the latent $\mathbf{Y}$ ones. Latent variables and observations are considered equal whenever the readings are fully observed, $\mathbf{y}_n = \mathbf{y}'_n$, and are considered different in those components where a MAR or a BDL occurs. Thus, a replacement step must be included in the algorithm, than accounts for these differences:

1. Fix the matrix of regression coefficients to a suitable value $\mathbf{B}^{(0)}$, e.g. intercept equal to a reasonable mean of the compositional response, slopes all equal to 0; fix suitable latent compositions $\mathbf{Z}^{(0)}$, e.g. as the data $\mathbf{Y}$ with irregular observations replaced in a sensible way;
2. simulate a random value $(\mathbf{\Sigma}^-)^{(k+1)}$ out of the distribution $\mathbf{\Sigma}^-|\mathbf{X}, \mathbf{Z}^{(k)}, \mathbf{B}^{(k)}$ of Eq. (11);
3. simulate a random value $\mathbf{B}^{(k+1)}$ out of the distribution $\mathbf{B}|\mathbf{X}, \mathbf{Z}^{(k)}, (\mathbf{\Sigma}^-)^{(k+1)}$ of Eq. (12);
4. for those n with missing values, simulate the latent readings $\mathbf{y}_n^{(k)}|\mathbf{y}_n^{'}, \mathbf{z}_n^{(k)}, \mathbf{P}$ from their associated latent errors $\mathbf{e}_n^{(k)}|\mathbf{e}_n^{'(k)}, \mathbf{P}$, which must follow a normal distribution in the same scheme as in step 4 of the preceding algorithm; this gives

$$\mathbf{e}_{mn} \sim \mathcal{N}_S\left((\mathbf{\Psi}_{mo} \cdot \mathbf{\Psi}_{oo}^{-1})\mathbf{e}_{on}^{(k)}, (\mathbf{\Psi}_{mm} - \mathbf{\Psi}_{mo} \cdot \mathbf{\Psi}_{oo}^{-1}\mathbf{\Psi}_{om})\right),$$

   where $\mathbf{\Psi} = -\mathbf{P}^{-1}/2$, and the blocks $(\mathbf{\Psi}_{mm}, \mathbf{\Psi}_{mo}; \mathbf{\Psi}_{om}, \mathbf{\Psi}_{oo})$ represent respectively the variances of the missing-missing, missing-observed, observed-missing and observed-observed subsets of variables of $\mathbf{y}_n^{'}$; finally, $\mathbf{y}_{mn}^{(k)} = \mathbf{z}_{mn}^{(k)} + \mathbf{e}_{mn}^{(k)}$;
5. simulate $\mathbf{Z}^{(k+1)}$ with the same Metropolis-Hastings algorithm as in section 3.2:
6. return to step 2 until the number of simulations is large enough.

5 Conclusions

Some sorts of missing values (missing completely at random and values below the detection limit) in compositional data sets can be dealt without need of a preliminary imputation. The key idea is to build a hierarchical model where these missing values are considered latent variables, to be estimated together with the model parameters. This contribution illustrates this concept with the problem of regression with compositional response. The proposed methodology requires a Bayesian framework and computationally intensive estimation techniques, through Markov Chain Monte Carlo methods (MCMC). Two models have been proposed.

The first model is the standard one in compositional regression, where the response is considered to follow an additive logistic normal distribution, with constant but unknown covariance and a mean built as an Aitchison affine linear combination of the explanatory variables. This is readily available to standard Bayesian multivariate regression techniques, as all parameters (regression coefficients, residual covariance and latent variables) have known conditional distributions (conjugate multivariate normal and Wishart models), thus amenable to Gibbs sampling techniques.

The second model builds upon the preceding one, including a layer of additive noise to the compositional response, i.e. assuming that observations are actually convolutions of a latent target composition plus a multivariate normal distribution. All observations are considered here different from their target composition, not just the missing ones. As with the first model, parameters (regression coefficients, residual covariances, additive noise covariances and latent compositions for *all* observations) can be obtained with standard MCMC techniques. In this case, the latent compositions can be sampled through an independent Metropolis-Hastings chain, while sampling all other parameters is possible with Gibss methods.

6 Acknowledgements

The authors gratefully acknowledge the support by the Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic).

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. London (UK): Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

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.

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

Gross, A. and R. Torres-Quevedo (1995). Estimating correlatons with missing data, a bayesian approach. *Psychometrika* (60), 341–354.

Gross, A. L. (2000). Bayesian interval estimation of multiple correlations with missing data: A gibbs sampling approach. *Multivariate Behavioral Research* (35), 201–227.

Mateu-Figueras, G., V. Pawlowsky-Glahn, and C. Barceló-Vidal (2003). Distributions on the simplex. In S. Thió-Henestrosa and J. A. Martín-Fernández (Eds.), *Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop*, Girona (E). Universitat de Girona, ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork2003/.

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

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics* (22), 1701–1762.

van den Boogaart, K., R. Tolosana-Delgado, and M. Bren (2011). The Compositional Meaning of a Detection Limit. In *Proceedings of the 4th International Workshop on Compositional Data Analysis (2011)*.

TrioScale: A new diagram for compositional data

MARK de ROOIJ[1] and PAUL EILERS[2]

[1]Institute of Psychology - Leiden University, The Netherlands, rooijm@fsw.leidenuniv.nl
[2]Department of Biostatistics, Erasmus University Medical Center, The Netherlands, p.eilers@erasmusmc.nl

1 Introduction

Compositional data (Aitchison, 1986) are data where the responses sum to a constant, typically 1 or 100. They occur in many areas of study. When there are three classes, the ternary diagram is very popular for displaying the data. It uses the fact that for any point within a triangle with equal sides, the sum of the perpendicular distances to the sides is constant.

The ternary diagram is attractive, but it has some drawbacks. First, it is hard to judge probabilities close to 0 or 1, because the scales are linear. It is desirable to expand the regions near the boundaries of the triangle, and in the corners, for a better view. Second, in our work on multinomial logistic regression models we found that axes for explanatory variables are heavily curved, making them hard to judge. In Section 3 more details will be provided.

The probabilistic ideal point model (Takane et.al, 1987; De Rooij, 2009) expresses probabilities in terms of distances between subjects/samples towards anchor points. Surprisingly, it turns out that this model holds the key to a transformation that eliminates the drawbacks of linear scales. It offers unlimited expansion of the boundary regions, as well as linear scales for covariates in a logistic model. This transformation, which we christened TrioScale can also be applied without any model.

2 TrioScale

Two types of compositional data can be distinguished: 1) Single set data where we have a matrix $\mathbf{F}$ with elements f_{ij} ($i = 1, \ldots, n$ and $j = 1, \ldots J$) from which we can define $p_{ij} = f_{ij}/f_{i+}$, i.e. the p_{ij} sum to one; and 2) Two set data where, besides the matrix $\mathbf{P}$, another set of variables is available for the samples i, i.e. $\mathbf{X}$ a $n \times p$ matrix of predictor variables. In the second case the f_{ij} are often (not necessarily) indicator matrices that indicate whether sample i is in category j or not. The first case can be turned into the second by defining $\mathbf{X}$ as a large identity matrix. In this paper the focus will be on compositional data with $J = 3$.

The ideal point model was primarily developed for modeling of a categorical response variable based on a set of predictor variables. Let $\mathbf{x}_i$ denote the values of the predictor variables for subject i, then

$$p_{ij} = p_j(\mathbf{x}_i) = \frac{\exp(-\delta_{ij})}{\sum_k \exp(-\delta_{ik})}$$

where δ_{ij} represents the squared Euclidean distance in two dimensions, that is

$$\delta_{ij} = (\eta_1(\mathbf{x}_i) - u_j)^2 + (\eta_2(\mathbf{x}_i) - v_j)^2$$

where $\eta_m(\mathbf{x}_i) = \alpha_m + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}_m$ is the coordinate of the position of subject/sample i on dimension m ($m = 1, 2$). The u_j and v_j are the coordinates on the first and second dimension, respectively, for *anchor point* j, representing the j-the response category. The α's will be collected in a vector $\mathbf{a} = [\alpha_1, \alpha_2]^{\mathsf{T}}$, the β's in a matrix $\mathbf{B} = [\boldsymbol{\beta}_1, \boldsymbol{\beta}_2]$. The η's are collected in the matrix $\mathbf{N} = [\eta_1, \eta_2]$. Finally, the coordinates of the anchor points are collected in the matrix $\mathbf{Z}$, that is $\mathbf{Z} = [\mathbf{u}, \mathbf{v}]$.

Having three categories in two-dimensional space we can place the anchor points at any position in the Euclidean sapce We will generally choose for an equilateral triangle. Having set the anchor points the parameters of the model ($\mathbf{a}$ and $\mathbf{B}$) can be estimated by maximizing the multinomial log likelihood

$$L = \sum_{i=1}^{n}\sum_{j=1}^{3} f_{ij} \log p_j(\mathbf{x}_i).$$

Often the interest focusses on the relationship between the predictor variables and the three classes. If so, variable axes need to be added to the diagram. The procedure follows that of biplots (Gower and Hand, 1996). For the j-th predictor variable define $\mathbf{m} = [\mathbf{0}, \zeta, \mathbf{0}]^\mathsf{T}$ a vector of zeros with on the j-th position the value ζ. Letting ζ vary within the range of observed values for the j-th predictor variable we obtain a series of points

$$\mathbf{n}_\zeta = \mathbf{a} + \mathbf{m}^\mathsf{T}\mathbf{B}$$

that form a linear trajectory in the two-dimensional space, the variable axis for variable j. Markers can be added to the variable axes by defining some interesting points ζ^* and add these. With many variables the display may become cluttered. In that case it is convenient to move the axes away from the anchor points and sample points. Blasius, Eilers, and Gower (2009) show for linear biplots how the axes can be shifted to the boundaries of the display, the same procedure can be used for our TrioScale.

Without predictor variables but with given p_{ij} a similar diagram can be made. Given a set of anchor points the position of a subject is completely determined. The coordinates $\mathbf{n} = [\eta_1, \eta_2]^\mathsf{T}$ for a subject with probabilities $\mathbf{p} = [p_{i1}, p_{i2}, p_{i3}]^\mathsf{T}$ can be computed as follows. The log odds, $\log \frac{p_{i1}}{p_{i2}}$, are a function of the distances, that is

$$\log\left(\frac{p_{i1}}{p_{i2}}\right) = \delta_{i2} - \delta_{i1}. \tag{1}$$

Using these the vector $\mathbf{r}$ can be defined with elements

$$\begin{aligned} r_1 &= \log(p_{i2}/p_{i1}) + v_2^2 - v_1^2 + u_2^2 - u_1^2, \\ r_2 &= \log(p_{i3}/p_{i1}) + v_3^2 - v_1^2 + u_3^2 - u_1^2. \end{aligned}$$

Furthermore, define the matrix

$$\mathbf{A} = \begin{bmatrix} u_2 - u_1, & v_2 - v_1 \\ u_3 - u_1, & v_3 - v_1 \end{bmatrix}.$$

The coordinates of the position of the subject is now given by

$$\mathbf{n} = \mathbf{A}^{-1}\mathbf{r}/2.$$

Samples can be included in TrioScale as dots. In cases where n is large the display may become to cluttered. Blasius, Eilers, and Gower (2009) propose smoothed density plots for the display of the samples.

For interpretation of the TrioScale diagram the log odds are of importance. The odds (derived above) are in favor of the closest category and are a simple difference of squared distances. Lines of constant odds are straight lines perpendicular to the line joining the two anchor points. The line exactly in the middle refers to the situation of equal odds, i.e. the distances are equal, so the log odds equals zero.

A second type of log-odds is $\log \frac{p_1}{1-p_1} = \log \frac{p_1}{p_2+p_3}$. Lines for constant $\log \frac{p_1}{1-p_1}$ are also lines for constant p_1. The probabilities were defined by

$$\begin{aligned} p_j &= \frac{\exp(-\delta_{ij})}{\sum_k \exp(-\delta_{ik})} \\ &= \frac{\exp(-\delta_{ij})}{\exp(-\delta_{i1}) + \exp(-\delta_{i2}) + \exp(-\delta_{i3})} \end{aligned}$$

When the coordinates of the anchor points are $(-1, 0)$, $(1, 0)$, and $(0, \sqrt{3})$ the equation for the line $p_1 = \pi$, a constant, is given by (see Appendix A)

$$\eta_2 = \frac{\log\left[\frac{(1-\pi)}{\pi}\exp(-2\eta_1 - 1) - \exp(2\eta_1 - 1)\right] + 3}{2\sqrt{3}},$$

similar equations can be derived for p_2 and p_3.

The circumcenter C of the triangle with coordinates $\mathbf{c} = [c_1, c_2]^\mathsf{T}$ formed by the anchor points is of special interest. The circumcenter is defined as the point which has equal Euclidean distance to the vertices of the triangle such that it refers to the case where the three probabilities are equal, $p_{i1} = p_{i2} = p_{i3} = 1/3$. The circumcenter is the point $[\eta_1, \eta_2]$ such that

$$\begin{aligned}(\eta_1 - u_1)^2 + (\eta_2 - v_1)^2 &= R^2\\ (\eta_1 - u_2)^2 + (\eta_2 - v_2)^2 &= R^2\\ (\eta_1 - u_3)^2 + (\eta_2 - v_3)^2 &= R^2\end{aligned}$$

where R is the radius of the circle passing though the three anchor points. The circumradius is a function of the length of the sides of the triangle (d)

$$R = \frac{d_1 d_2 d_3}{4\Delta},$$

where Δ is the area of the triangle, see De Rooij and Gower (2003).

In principle the anchor points can be positioned anywhere in the two dimensional space as long as they span two dimensions, some choices are however more natural than others. One natural choice is to place the anchor points as the vertices of an equilateral triangle with the origin as circumcenter and point of gravity. Compared to linear ternary diagrams the positions of the samples are not necessarily within the triangle. However, by making the triangle larger it is always possible that the samples fall within the boundaries of the triangle.

A second natural choice is to position the anchor points

$$\mathbf{Z}_1 = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

This set of coordinates has the advantage that the regression weights from a multinomial logistic regression are obtained up to a factor 2. Any two sets of anchor points are naturally related and can be written as a function of each other. Therefore, write the ideal points as

$$\mathbf{N} = \mathbf{1}\mathbf{a}^\mathsf{T} + \mathbf{X}\mathbf{B}.$$

With two equivalent solutions we have $\mathbf{a}_1$ and $\mathbf{a}_2$, $\mathbf{B}_1$ and $\mathbf{B}_2$, and $\mathbf{Z}_1$ and $\mathbf{Z}_2$. One set of anchor points can be written as the following linear function of the other set

$$\mathbf{Z}_2 = \mathbf{1}\mathbf{v}^\mathsf{T} + \mathbf{Z}_1\mathbf{T}.$$

Applying the inverse operation on the regression weights we obtain

$$\mathbf{B}_2 = \mathbf{B}_1(\mathbf{T}^{-1})^\mathsf{T}$$

Finally, for the intercepts $\mathbf{a}_2$ and $\mathbf{a}_1$ we have the following relation

$$(\mathbf{a}_2 - \mathbf{c}^2) = (\mathbf{a}_1 - \mathbf{c}^1)(\mathbf{T}^{-1})^\mathsf{T}$$

where $\mathbf{c}^2$ is the coordinate vector of the circumcenter of the second set and $\mathbf{c}^1$ of the first set. Not the vector $\mathbf{v}$ is of importance here but the circumcenters of the two triangles.

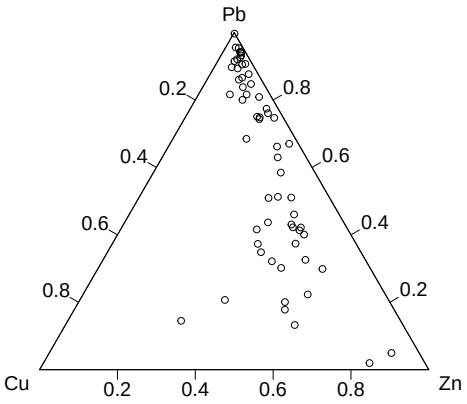


Figure 1: Ternary diagram with probabilities close to zero and one.

3 Applications

3.1 Simulated data with large probabilities

In the introduction we alluded that the standard ternary diagram is difficult to interpret with probabilities close to zero or one. For an example we use data that were simulated using the R package *compositions* (Van den Boogaart, Tolosana, and Bren, 2012). Figure 1 shows the standard linear diagram where the fractions of copper (Cu) are very small such that the samples are all near to the

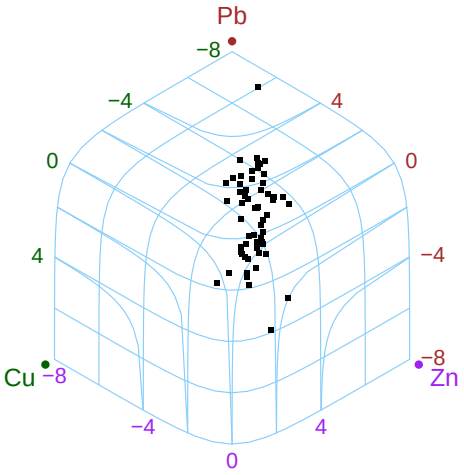


Figure 2: TrioScale diagram with probabilities close to zero and one. The light blue lines represent constant levels of the log-odds. Their values are marked at the boundary in the colors that correspond to those of the anchor points.

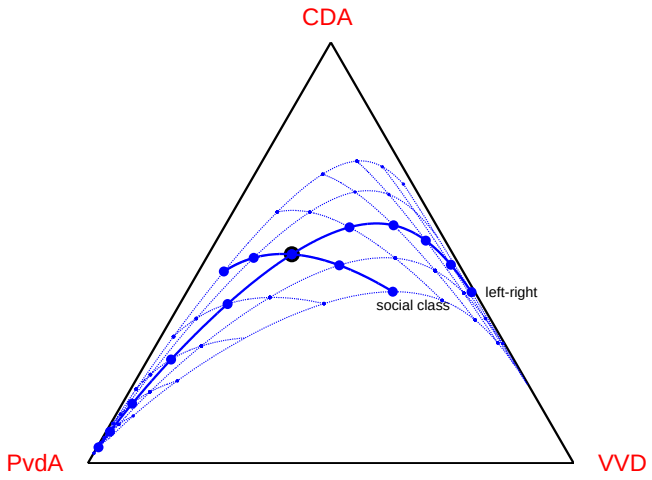


Figure 3: Ternary diagram with variable axes.

boundary and close to the axis for lead (Pb).

Figure 2 shows the TrioScale diagram for the same data. The points are not so cluttered anymore. Three sets of level lines for constant values of the log-odds $\log(p_j/(1-p_j))$ have been added. For p_2 they consist of a rounded part, symmetric about the vertical through the top anchor point, that connects two (asymptotically) straight lines which span an angle of 120 degrees. The distance between level lines is linearly proportional to a to the change in log-odds. It can be extended as far as needed, to very high and very low log-odds. For p_1 (p_3) the system of level lines is rotated 120 (-120) degrees clockwise.

The values of the log-odds levels are provided in the display. To obtain some feeling of these values, with a log odds of -4 the probability is 0.018, with -2 the probability is 0.12, with 0 the probability is 0.50, with +2 the probability is 0.88, and with +4 the probability is 0.98.

3.2 Empirical data with covariates

If covariates are available, a logistic regression model can be fitted. In linear ternary diagrams axes for the covariates can be included. However, such axes are smooth but strongly nonlinear curves such that with multiple predictor variables the interpretation in terms of conditional effects is awkward. As an illustration we consider data from the Dutch parliamentary election (DPES) data of 2006. In this survey people were asked their political vote in the 2006 election. Our attention focusses to the subjects that voted on either the Labor Party (PvdA), the Christian Democratic party (CDA), and the Conservative liberals (VVD). These three parties are traditionally the three largest in the Netherlands. Several background variables are available, for simplicity we focus on two:

- Left-Right Self-rating: An assessment of one's own position on the left-right political continuum (11 point scale), with score -5 indicating left and +5 indicating right.
- Social Class: An assessment of one's own social class with scores -2 working class; -1 upper working class; 0 middle class; 1 upper middle class; 2 upper class.

A multinomial logistic regression model was fitted. Figure 3 shows the estimated probabilities and the variable axes, which are strongly nonlinear curves. This nonlinearity makes it difficult to

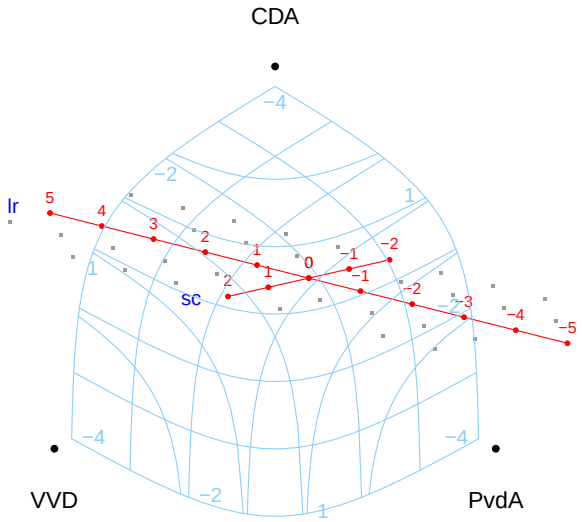


Figure 4: The TrioScale diagram for the Dutch Parliamentary Election Studies.

understand conditional effects. Therefore, the complete grid representing all 55 possible patterns is added, but it is of little help: e.g. it is hard to find the position for a subject with scores -4 and 1. With more than three predictor variables the conditional effects are impossible to understand from the display.

Figure 4 shows the TrioScale representation of the same data. The variable axes are linear now and scale markers are easily added. Positions of subjects can be simply obtained by completing parallelograms, i.e. for a subject with values 4 on left-right and -1 on social class simply go 4 units on the variable *lr* and -1 on *sc* and complete the parallelogram. This procedure can be easily generalized for multiple predictor variables.

Details of the similarities between ideal point models, multinomial logistic regression and log-ration analysis are discussed in Appendix B.

3.3 Hardy-Weinberg equilibrium

The Hardy-Weinberg equilibrium states that genotype frequencies in a population will remain constant from generation to generation in the absence of evolutionary influences and under random mating (Wikipedia). Each genotype is composed of an unordered pair of two alleles, say A and a and thus can be denoted as AA, Aa or aa. Their respective probabilities are p^2, $2p(1-p)$ and $(1-p)^2$, if p is the probability of A. In a standard ternary diagram, such an equilibrium is represented by a smooth curve from the AA point towards the aa point. In TrioScale such an equilibrium results in a straight horizontal line (see Appendix C). To illustrate this, data were generated using the HardyWeinberg package in R (Graffelman, 2012). Figure 5 shows the TrioScale solution for the generated data. The samples all lie on a straight line.

4 Conclusion

We proposed, TrioScale, an alternative to the standard ternary diagram, with several important advantages: probabilities close to zero or one are better visualized and variable axes become linear and additive.

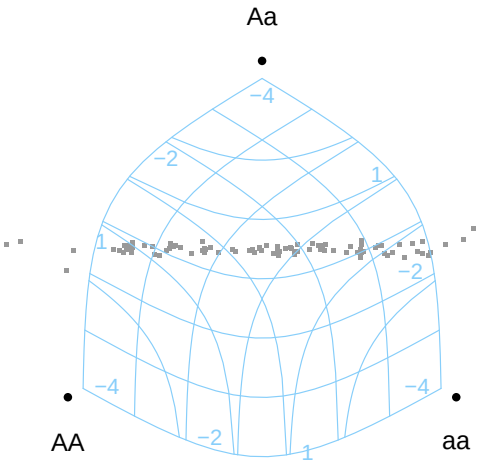


Figure 5: TrioScale diagram for simulated Hardy Weinberg data.

For the analysis of data and the graphical representation a set of R-scripts were written by the authors. The plan is to turn these into a R-package. For the moment, the scripts can be obtained from the authors.

References

Agresti, A. (2002). *Categorical Data Analysis*. John Wiley& Sons, Hoboken, NJ (USA).

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.

Blasius, J. Eilers, P, and Gower, J.. (2009). Better Biplots *Computational Statistics and Data Analysis 53*, 3145–3158.

De Rooij, M. (2009). Ideal point discriminant analysis with a special emphasis on visualization *Psychometrika 74*, 317–330.

De Rooij, M. and Gower, J. (2003). The geometry of triadic distances *Journal of Classification 20*, 181–220.

Gower, J.C. and Hand, D.J. (1996). *Biplots*. Chapman & Hall, London (UK).

Graffelman, J. (2012). HardyWeinberg: Graphical tests for Hardy Weinberg Equilibrium R package version 1.4.1 http://CRAN.R-project.org/package=HardyWeinberg

Takane, T., Bozdogan, H. and Shibayama, T. (1987). Ideal point discriminant analysis *Psychometrika 52*, 371–392.

Van den Boogaart, K.G. and Tolosana, R. and Bren, M (2012). compositions: Compositional Data Analysis. R package version 1.20-1. http://CRAN.R-project.org/package=compositions

Appendix A: Equations for constant p_1

The coordinates of the three anchor points are again $(-1,0)$, $(1,0)$ and $(0,\sqrt{3})$. We will find the equation for η_1, η_2 for $p_1 = \pi$, a constant $(0 < \pi < 1)$. Lets start with the definition of the probability and working out the terms using the coordinates of the anchor points

$$\begin{aligned} p_1 = \pi &= \frac{\exp\left(2\eta_1 u_1 - u_1^2 + 2\eta_2 v_j - v_j^2\right)}{\exp\left(2\eta_1 u_1 - u_1^2 + 2\eta_2 v_j - v_j^2\right) + \exp\left(2\eta_1 u_2 - u_2^2 + 2\eta_2 v_2 - v_2^2\right) + \exp\left(2\eta_1 u_3 - u_3^2 + 2\eta_2 v_3 - v_3^2\right)} \\ &= \frac{\exp\left(-2\eta_1 - 1\right)}{\exp\left(-2\eta_1 - 1\right) + \exp\left(2\eta_1 - 1\right) + \exp\left(2\sqrt{3}\eta_2 - 3\right)} \\ &= \frac{\exp\left(-2\eta_1 - 1\right)}{\exp\left(-2\eta_1 - 1\right) + \exp\left(2\eta_1 - 1\right) + \exp\left(2\sqrt{3}\eta_2 - 3\right)}. \end{aligned}$$

By rewriting we obtain

$$\pi\left[\exp\left(-2\eta_1 - 1\right) + \exp\left(2\eta_1 - 1\right) + \exp\left(2\sqrt{3}\eta_2 - 3\right)\right] = \exp\left(-2\eta_1 - 1\right),$$

and by shuffling the terms

$$\pi \exp\left(2\sqrt{3}\eta_2 - 3\right) = \exp\left(-2\eta_1 - 1\right) - \pi\exp\left(-2\eta_1 - 1\right) - \pi\exp\left(2\eta_1 - 1\right)].$$

Finally, rewriting the equation and simplifying we have

$$\pi \exp\left(2\sqrt{3}\eta_2 - 3\right) = (1-\pi)\exp\left(-2\eta_1 - 1\right) - \pi\exp\left(2\eta_1 - 1\right),$$

$$\exp\left(2\sqrt{3}\eta_2 - 3\right) = \frac{(1-\pi)}{\pi}\exp\left(-2\eta_1 - 1\right) - \exp\left(2\eta_1 - 1\right),$$

$$\left(2\sqrt{3}\eta_2 - 3\right) = \log\left[\frac{(1-\pi)}{\pi}\exp\left(-2\eta_1 - 1\right) - \exp\left(2\eta_1 - 1\right)\right],$$

$$\eta_2 = \frac{\log\left[\frac{(1-\pi)}{\pi}\exp\left(-2\eta_1 - 1\right) - \exp\left(2\eta_1 - 1\right)\right] + 3}{2\sqrt{3}}.$$

Similar equations can be obtained for p_2 and p_3. Here we used a fixed set of anchor points, transformation to another set of anchor points can be obtained from the equations described in Section 2.

Appendix B: Multinomial logistic regression and log-ratio analysis

The ideal point model and TrioScale have relationships to multinomial logistic regression and log ratio analyses. In *multinomial logistic regression* a baseline category is chosen, we take the first category. Then Multinomial logistic regression is defined by the system of equations

$$p_j(\mathbf{x}_i) = \frac{\exp(\alpha_j + \mathbf{x}_i^{\mathsf{T}}\beta_j)}{1 + \sum_k \exp(\alpha_k + \mathbf{x}_i^{\mathsf{T}}\beta_k)},$$

with $\alpha_1 = 0$ and $\beta_1 = 0$. There is an equivalent log-odds representation given by

$$\log(p_j(\mathbf{x}_i)/p_1(\mathbf{x}_i)) = \alpha_j + \mathbf{x}_i^{\mathsf{T}}\beta_j,\ j = 2,3.$$

The multinomial logistic regression model is also estimated by optimizing the log likelihood function given above. For more details see Agresti (2002). De Rooij (2009) showed that the model defined above is actually equal to a multinomial logistic regression. Starting with the ideal point model rewriting it

$$\begin{aligned} p_{ij} = p_j(\mathbf{x}_i) &= \frac{\exp\left(-\delta_{ij}\right)}{\sum_k \exp\left(-\delta_{ik}\right)} \\ &= \frac{\exp\left(-\left(\eta_1(\mathbf{x}_i) - u_j\right)^2 - \left(\eta_2(\mathbf{x}_i) - v_j\right)^2\right)}{\sum_k \exp\left(-\left(\eta_1(\mathbf{x}_i) - u_k\right)^2 - \left(\eta_2(\mathbf{x}_i) - v_k\right)^2\right)} \\ &= \frac{\exp\left(2\eta_1(\mathbf{x}_i)u_j - u_j^2 + 2\eta_2(\mathbf{x}_i)v_j - v_j^2\right)}{\sum_k \exp\left(2\eta_1(\mathbf{x}_i)u_k - u_k^2 + 2\eta_2(\mathbf{x}_i)v_k - v_k^2\right)}. \end{aligned}$$

Defining $u_1 = 0, u_2 = 1, u_3 = 0$ and $v_1 = 0, v_2 = 0, v_3 = 1$ the model probabilities become

$$\begin{aligned} p_{i1} &= \frac{\exp\left(2\eta_1(\mathbf{x}_i)0 - 0^2 + 2\eta_2(\mathbf{x}_i)0 - 0^2\right)}{\sum_k \exp\left(2\eta_1(\mathbf{x}_i)u_k - u_k^2 + 2\eta_2(\mathbf{x}_i)v_k - v_k^2\right)} \\ &= \frac{\exp(0)}{\sum_k \exp\left(2\eta_1(\mathbf{x}_i)u_k - u_k^2 + 2\eta_2(\mathbf{x}_i)v_k - v_k^2\right)} \\ p_{i2} &= \frac{\exp\left(2\eta_1(\mathbf{x}_i)1 - 1^2 + 2\eta_2(\mathbf{x}_i)0 - 0^2\right)}{\sum_k \exp\left(2\eta_1(\mathbf{x}_i)u_k - u_k^2 + 2\eta_2(\mathbf{x}_i)v_k - v_k^2\right)} \\ &= \frac{\exp\left(2\eta_1(\mathbf{x}_i) - 1\right)}{\sum_k \exp\left(2\eta_1(\mathbf{x}_i)u_k - u_k^2 + 2\eta_2(\mathbf{x}_i)v_k - v_k^2\right)} \\ p_{i3} &= \frac{\exp\left(2\eta_1(\mathbf{x}_i)0 - 0^2 + 2\eta_2(\mathbf{x}_i)1 - 1^2\right)}{\sum_k \exp\left(2\eta_1(\mathbf{x}_i)u_k - u_k^2 + 2\eta_2(\mathbf{x}_i)v_k - v_k^2\right)} \\ &= \frac{\exp\left(2\eta_2(\mathbf{x}_i) - 1\right)}{\sum_k \exp\left(2\eta_1(\mathbf{x}_i)u_k - u_k^2 + 2\eta_2(\mathbf{x}_i)v_k - v_k^2\right)}. \end{aligned}$$

Now, using $\eta_m(\mathbf{x}_i) = \alpha_m + \mathbf{x}_i^{\mathsf{T}}\boldsymbol{\beta}_m$, we see that the -1 will be included in the intercept and the factor 2 will be captured in the regression weights. The same probabilities are obtained using both models, thus the same log likelihood is achieved by both modeling tools. The ideal point model defined above can thus be seen a graphical tool to visually aid the interpretation of the multinomial logistic regression.

Log ratio analysis analysis starts with the proportions p_{ij} and builds log ratios from them, i.e. $\log(p_{i1}/p_{i3})$ and $\log(p_{i2}/p_{i3})$. This is essentially the same as our approach. However, the log ratio approach is often used with more categories in the response variable. In that case, first the log ratios are computed and second a principal components analysis (or another multivariate technique) is used to reduce the dimension of the solution. Notice, that in the case of zero/one response data, i.e. the log ratio approach fails since the log ratios cannot be computed.

Appendix C: Hardy Weinberg Equilibrium

In the case of Hardy-Weinberg equilibrium we have

$$\begin{aligned} p(AA) &= \pi^2 \\ p(aa) &= (1-\pi)^2 \\ p(Aa) &= 2\pi(1-\pi). \end{aligned}$$

The following log odds are defined:

$$\theta_1 = \log\left(\frac{p(AA)}{p(Aa)}\right)$$
$$\theta_2 = \log\left(\frac{p(aa)}{p(Aa)}\right)$$

Let us make a log odds display with AA on position $(-1,0)$, aa on $(1,0)$ and Aa on $(0,\sqrt{3})$. We are going to look for positions in the two dimensional space with coordinates η_1 and η_2 on the horizontal and vertical axis respectively with probabilities given by the Hardy-Weinberg equilibrium. Therefore, we have that

$$\begin{aligned}\theta_1 &= \delta_{Aa}^2 - \delta_{AA}^2 \\ &= (\eta_1-0)^2 + (\eta_2-\sqrt{3})^2 - [(\eta_1+1)^2 - (\eta_2-0)^2] \\ &= \eta_1^2+\eta_2^2+3-2\sqrt{3}\eta_2 - [\eta_1^2+1+2\eta_1] - \eta_2^2 \\ &= 3-2\sqrt{3}\eta_2 - 1 - 2\eta_1,\end{aligned}$$

and

$$\begin{aligned}\theta_2 &= \delta_{Aa}^2 - \delta_{aa}^2 \\ &= (\eta_1-0)^2 + (\eta_2-\sqrt{3})^2 - [(\eta_1-1)^2 - (\eta_2-0)^2] \\ &= \eta_1^2+\eta_2^2+3-2\sqrt{3}\eta_2 - [\eta_1^2+1-2\eta_1] - \eta_2^2 \\ &= 3-2\sqrt{3}\eta_2 - 1 + 2\eta_1,\end{aligned}$$

from which follows that

$$\begin{aligned}-2\eta_1 &= -3+2\sqrt{3}\eta_2+1+\theta_1 \\ 2\eta_1 &= -3+2\sqrt{3}\eta_2+1+\theta_2.\end{aligned}$$

Dividing by 2 and filling the first into the second we obtain

$$\eta_2 = \frac{3-(\theta_1/2+\theta_2/2)}{2\sqrt{3}}.$$

Finally we have

$$\begin{aligned}\theta_1+\theta_2 &= \log\left(\frac{\pi^2}{2\pi(1-\pi)}\right) + \log\left(\frac{(1-\pi)^2}{2\pi(1-\pi)}\right) \\ &= \log(\pi^2) - \log(2\pi(1-\pi)) + \log(\pi^2) - \log(2\pi(1-\pi)) \\ &= \log(\pi) + \log(\pi) - (\log(2)+\log(\pi)+\log(1-\pi)) \\ &\quad + \log(1-\pi) + \log(1-\pi) - (\log(2)+\log(\pi)+\log(1-\pi)) \\ &= -2\log(2).\end{aligned}$$

Inserting this into 2 gives

$$\begin{aligned}\eta_2 &= \frac{3-(\theta_1/2+\theta_2/2)}{2\sqrt{3}} \\ &= \frac{3-(\theta_1+\theta_2)/2}{2\sqrt{3}} \\ &= \frac{3+\log(2)}{2\sqrt{3}},\end{aligned}$$

so η_2 is a constant and the Hardy Weinberg equilibrium leads to a horizontal line in our display.

Testing compositional association

J. J. EGOZCUE[1], D. LOVELL[2] and V. PAWLOWSKY-GLAHN[3]

[1]Dept. Applied Mathematics III, U. Politécnica de Catalunya, Barcelona, Spain, juan.jose.egozcue@upc.edu
[2] CSIRO Mathematics, Informatics, and Statistics, Canberra, Australia
[3] Dept. Informatics and Applied Mathematics, U. de Girona, Spain

Abstract

Proportionality is a meaningful measure of association for parts of compositions: when two parts are perfectly proportional, the variance of their log-ratio is zero. However, when they are not perfectly proportional, it is not immediately clear how to interpret the positive variance of their log-ratio. Two methods for testing *compositional association* of two parts are presented. They check whether an observed positive variance of a log-ratio is significantly different from what one would see if the parts were approximately proportional. As a first step a normalisation of the variation array is proposed. This involves computing the proportion of the total variance of the sample explained by the variance of a simple log-ratio, making the normalized values comparable to other log-ratios of the same composition. Testing techniques are more involved. The main difficulty comes from the fact that the natural null hypothesis is zero variance of the log-ratio. Under this null hypothesis, the log-ratio has no variability and *any* sample variability implies rejection of the null hypothesis. Instead, we look to compare variability of the simple log-ratio against other sources of variability. Two approaches have been successfully used. The first one considers the linear regression

$$b_{i1} = \beta_0 + \sum_{j=1}^{D} \beta_j \; b_{ij} + \epsilon_i \; , \; i = 1, 2, \ldots, n \; ,$$

where b_{i1} is the sample balances involving the two reference parts, and b_{ij} are the sample balances orthogonal to b_{i1}. A non-significant F-test suggests that the balance b_1 is approximately constant up to residuals ϵ_i. The main shortcomings of this procedure are that the source of variability is restricted to the sample, and the null hypothesis is only a necessary, but not sufficient, condition for proportionality of the reference parts. The second approach considered is the regression

$$\mathrm{clr}_1(\mathbf{x}_i) = \beta_0 + \beta_1 \mathrm{clr}_2(\mathbf{x}_i) + \epsilon_i \; , \; i = 1, 2, \ldots, n \; ,$$

where $\mathbf{x}_i$ are the sample compositions and subscripts 1, 2 refer to reference parts. It can be tested for $\beta_1 = 1$ using a $\mathtt{t}$-statistics, due to the fact that $\beta_1 = 1$ implies proportionality of parts 1, 2. The fact that this regression is not symmetric in the two parts, and total least squares do not work in this case, leads to consider the symmetrical regressions and consider the probit-average of the two p-values. Again, non-significant p-values are expected for well associated parts.

The normalised variation array, together with the two proposed p-values, appears to be an exploratory tool assessing compositional association, able to replace the misleading correlation matrix of raw compositional data.

1 Introduction

Association between parts of a composition can be conceived in several ways. However, simplicity and interpretability should be the main ideas when selecting measures of association. Historically, the Pearson correlation has been the main association measure in multivariate analysis. It is simple, as it relates only two variables of a random vector; it concerns only linear transformation in $\mathbb{R}^n$, i.e. change of scale plus a shift. Interpretation relies on the linear regression ideas, which in turn are related to the geometry of $\mathbb{R}^n$, where covariance appears as a Euclidean inner product in the space of samples. All these desirable properties fail when Pearson correlation is applied to study association between parts of a composition. Spurious correlation makes Pearson correlation useless and confusing when applied to the parts of a composition (Pearson, 1897; Aitchison, 1986; Lovell et al., 2013). As a consequence, most real multivariate techniques break down for compositions treated as multivariate vectors.

The principle of working in coordinates (Mateu-Figueras et al., 2011) gives a way out: it advocates that compositions be represented in coordinates of the simplex, mainly orthonormal ones (also called

ilr-coordinates), and then treats coordinates as real multivariate vectors. In ilr-coordinates, Pearson correlation takes again the central role it has in real multivariate statistics. However, interpretability can be demanding when trying to interpret a correlation between two ilr-coordinates, as they involve, at least, three parts of the composition.

Experience suggests that compositional data analysts tend to interpret single parts of a composition even knowing that this is impossible. The association between two parts is then the simplest possible association that can be consistently analysed. This fact was recognized from the beginning of compositional analysis, when the variation matrix was introduced as a way to represent the variability of a random composition (Aitchison, 1986).

Geochemistry lends two examples of compositional association that can be assumed as paradigmatic. The first one is the chemical equilibrium, in which some chemical species A_i, $i = 1, 2, \ldots, n_A$ react to give species B_j, $j = 1, 2, \ldots, n_B$ and vice versa. If concentrations are given in ppm of mass, the equilibrium is normally symbolized as

$$\alpha_1 A_1 + \alpha_2 A_2 + \cdots + \alpha_{n_A} A_{n_A} \rightleftharpoons \beta_1 B_1 + \beta_2 B_2 + \cdots + \beta_{n_B} B_{n_B} \ , \tag{1}$$

where the α coefficients equal the β coefficients due to the conservation of mass. When the reaction attains equilibrium and the rate of reaction of A's to B's is equal to the reverse reaction, the log-ratio

$$\log \frac{\prod_{i=1}^{n_A} A_i^{\alpha_i}}{\prod_{j=1}^{n_B} B_j^{\beta_j}} \tag{2}$$

is assumed to be constant, as the numerator and denominator are proportional to the rates of reaction from one side of the equilibrium to the other. Moreover, as $\sum \alpha_i = \sum \beta_j$, the log-ratio (2) is a log-contrast. Therefore, the chemical equilibrium, expressing association between the group of species A and the group of species B, appears as a constant log-contrast.

Stoichiometry is a second kind of association between two groups of chemical species. Mineral crystals satisfy some chemo-electrical equilibrium conditions controlling the proportions of the different chemical species involved. Some species are electrically equivalent and can substitute each other. A stoichiometric relation can be expressed as

$$\alpha_1 A_1 + \alpha_2 A_2 + \cdots + \alpha_{n_A} A_{n_A} = \beta_1 B_1 + \beta_2 B_2 + \cdots + \beta_{n_B} B_{n_B} \ , \tag{3}$$

where the species within each group, the A's and the B's, are more or less exchangeable within the group, taking into account their abundances in the growing process and the physical structure of the crystal mesh. Note the difference between a chemical equilibrium equation (1) and a stoichiometric rule (3). In the first one, the plus signs and the $\rightleftharpoons$ sign are symbols which imply the form of the association: the constant log-contrast (2). Alternatively, the stoichiometric relation is described mathematically with (3), i.e. the sums are actual sums and equality is the standard equality of real numbers. From the compositional point of view, the stoichiometric relation (3) is a non-linear equation, as amalgamation and real product by scalars are not simple operations in the simplex. This is a good reason for not using stoichiometric relations as association measures between parts of a composition.

The fact that a constant log-contrast in a composition determines a hyperplane in the simplex suggests that constant log-contrasts can be viewed as the perfect (linear) association between groups of parts. The simplest log-contrast is a log-ratio involving only two parts. Consequently, two parts of a composition are perfectly associated when their log-ratio is constant or, equivalently, when the two parts are exactly proportional. However, two parts are seldom exactly proportional across a sample and the approximate proportionality should be quantified.

For illustration purposes, an example will be used along this contribution. It consists of 29 samples of ground water for which the concentrations of major anions and cations are given (mg/l). The samples were published in Moeller et al. (2008). Data were obtained on a study of salinization of deep groundwater in the North German basin. The major anions and cations considered here are Na, K, Mg, Ca, Cl, SO_4, HCO_3, where the electric charge of each ion has been suppressed for simplicity. Taking into account that the study was on salinization, one expects a good association between Na and Cl as coming from dissolved salt from deep brines. This example has been used previously for

illustration of a compositional exploratory and regression analysis by Egozcue and Pawlowsky-Glahn (2011).

This contribution aims at discussing measures of proportionality, i.e. compositional association. A normalization of the variation array is a first useful approach presented in Section 2. Sections 3 and 4 suggest two testing procedures linked to compositional association of two parts.

2 Normalization of a variation array

To date, the variation matrix introduced by Aitchison (1986) has been the main tool measuring association of two parts in compositional data analysis. For a D-part random composition $X = (X_1, X_2, \ldots, X_D)$, consider an n-sample $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$, $i = 1, 2, \ldots, n$. The sample variation matrix $\mathbf{V}$, is a (D, D) array with j, k entry equal to

$$\mathrm{Var}\left[\log \frac{X_j}{X_k}\right] = \frac{1}{n}\sum_{i=1}^{n} \log^2\left(\frac{x_{ij}}{x_{ik}}\right) - \left[\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{x_{ij}}{x_{ik}}\right)\right]^2 .$$

The elementary properties of $\mathbf{V}$ are well-known: (a) it is symmetric; (b) its diagonal elements are zero, and non-negative entries out off the diagonal; (c) the sum of all entries is twice the sample total variance.

When an element of $\mathbf{V}$ is close to zero, it means that the ratio X_j/X_k has a small variability around a constant value, i.e. X_j, X_k are almost proportional. This fact has been mentioned in different contexts, and used explicitly to build *principal balances* by Pawlowsky-Glahn et al. (2011, 2013) using hierarchical clustering of parts.

However, there is a lack of scale in the elements of the variation matrix, and the meaning of *close to zero* or *small* should be considered with caution. A first normalization consists in dividing the variation matrix by the total sum of the entries, which is $2D \cdot \mathrm{TotVar}$. This normalization presents the variation matrix $\mathbf{V}$ as a $D(D-1)$ composition, but the result is still hard to visualize. A second, more useful, normalization is based on the idea that "completely non-proportional" parts of a D-composition correspond to uniform values off the diagonal. In this case, each of these entries would have the value $2D \cdot \mathrm{TotVar}/(D(D-1)) = 2\mathrm{TotVar}/(D-1)$. Dividing each off diagonal entry by this value yields a perturbation matrix for complete non-proportionality; its sum adds to the number of off diagonal entries, and is thus equal to $D(D-1)$. The resulting normalised variation matrix is

$$\tilde{\mathbf{V}} = \frac{D-1}{2\mathrm{TotVar}}\mathbf{V} .$$

An entry of $\tilde{\mathbf{V}}$ greater than one means that the corresponding couple of parts is less proportional than the log-ratio variance that would be observed in a completely non-proportional composition. Values less than one indicate association; the smaller is the entry, the more associated is the couple of parts. These aspects are illustrated in Table 1.

Table 1: Variation matrix of Moeller et al. (2008) data. Upper triangle is the original variation matrix; the lower triangle is the normalised version. Traces of association are presented in blue. Larger associations (lower values) are in bold face; if less than 0.30 in blue, if less than 0.20 in violet. Non-proportionality in red. Total variance is 7.583.

	Na	K	Mg	Ca	Cl	SO_4	HCO_3
Na	0.000	0.647	0.736	1.262	0.291	1.549	5.962
K	**0.256**	0.000	0.293	0.876	1.344	2.219	4.653
Mg	**0.291**	**0.116**	0.000	0.520	1.408	2.261	4.475
Ca	0.499	0.347	**0.206**	0.000	2.147	2.547	4.382
Cl	**0.115**	0.532	0.557	0.849	0.000	1.884	7.714
SO_4	0.613	0.878	0.895	1.008	0.745	0.000	5.910
HCO_3	2.359	1.841	1.771	1.734	3.052	2.338	0.000

Table 1 shows that HCO_3 is highly non-proportional to all other elements. The cation SO_4 is also non-proportional to the other elements. In contrast, Na appears associated with all anions,

particularly Cl. An hypothetical association between K and Mg can be geologically interpreted. The possible association of Na and Cl can be relevant in a salinization study. However, the values of association of the two couples of elements are not very low, and an analyst can ask for checking these associations to be true, up to the available data. This means that, being the normalized variation matrix straightforward to interpret, some kind of statistical testing is required.

3 Testing proportionality

Searching for an hypothesis testing related to compositional association—i.e., proportionality of two parts—has several problems. The main one is that under the natural null hypothesis—*i.e., that the two parts are proportional*—the variability of most adequate statistics is zero. It is similar to the situation in which the null hypothesis on a random variable is that it has zero variance: any variability of the statistics leads to the immediate rejection of the null hypothesis. This is the reason why, in a regression analysis, the null hypothesis *the multiple correlation coefficient is 1* is never stated. Accordingly, a proper hypothesis testing that *the two parts are proportional* cannot be used. In its place, hypotheses related, but not equivalent, to proportionality are tested.

Assuming that perfect compositional association consists of proportionality of two parts, say X_1, X_2, it can be formulated as $X_1 = aX_2$ for any positive constant a. If some departures of exact proportionality are assumed, a linear regression model could be obtained by adding residuals to the condition of proportionality. This strategy leads to non-scaled residuals and a non-scale invariant model. Next step consists in writing the proportionality condition after taking logarithms, i.e. $\log X_1 = \log a + \log X_2$. Adding residuals to this log-condition, the corresponding regression model is more consistent, but still the residuals are not scaled according to the Aitchison metric of the simplex. A simple way to overcome this obstacle is to transform the logarithms to clr components subtracting the log of a geometric mean of some parts. If the geometric mean includes only X_1 and X_2 the model becomes degenerate. Alternatively, the geometric mean can include all the parts of the composition under study. Consequently, the new model

$$\mathrm{clr}_1(\mathbf{x}_i) = \beta_0 + \beta_1 \mathrm{clr}_2(\mathbf{x}_i) + \epsilon_i \quad , \quad i = 1, 2, \ldots, n \ , \tag{4}$$

is proposed. To check the relationship with the proportionality condition, this can be rewritten as

$$\log(x_{i1}) = \beta_0 + (1 - \beta_1)\mathrm{g_m}(\mathbf{x}_i) + \beta_1 \log(x_{i2}) + \epsilon_i \quad , \quad i = 1, 2, \ldots, n \ ,$$

which is reduced to exact proportionality with zero residuals for $\beta_0 = \log a$ and $\beta_1 = 1$. The advantage of this model is that residuals are in the clr-plane, in the direction of $\mathrm{clr}_1(\mathbf{x}_i)$, and their norm corresponds to the Aitchison metric.

The hypothesis $H_0 : \ \beta_1 = 1$, against $H_1 : \ \beta_1 \neq 1$, can be tested on the model (4). H_0 can be tested in regression analysis under normality hypothesis of the residuals. The test can be, equivalently, stated in terms of `t`-student statistics, χ^2-statistics, and F-statistics. Adopting the `t`-student strategy, and denoting $z = \mathrm{clr}_1(\mathbf{x})$, $y = \mathrm{clr}_2(\mathbf{x})$, the statistic is

$$T_1 = \frac{\widehat{\beta_1} - 1}{S_{\widehat{\beta_1}}} = \frac{\widehat{\beta_1} - 1}{(\sqrt{n/(n-2)})S_\epsilon} \sqrt{nS_y^2} \ ,$$

where $S_{\widehat{\beta_1}}$ is the standard deviation of the estimator of β_1 and

$$\widehat{\beta_0} = \bar{z} - \widehat{\beta_1}\bar{y} \ , \ \widehat{\beta_1} = \frac{S_{zy}}{S_y^2} \ , \ S_\epsilon^2 = \frac{1}{n}\sum_i (z_i - \widehat{\beta_0} - \widehat{\beta_1} y_i)^2 \ ,$$

$$\bar{z} = \frac{1}{n}\sum_i z_i \ , \ \bar{y} = \frac{1}{n}\sum_i y_i \ .$$

The statistic has distribution $T_1 \sim t_{n-2}$. The two-tails of the distribution allow the computation of the corresponding p-value. This test on proportionality based on the correlation of clr-components is called clr-slope test (clrST).

The main problem with this test is that it is not symmetric in $\mathrm{clr}_1(\mathbf{x})$ and $\mathrm{clr}_2(\mathbf{x})$, as the residuals are measured in different clr directions. Therefore, testing the model (4) after swapping $\mathrm{clr}_1(\mathbf{x})$ and $\mathrm{clr}_2(\mathbf{x})$, i.e. permutation of z and y, results in a different statistics T_2 and a different p-value. The two p-values, say p_1 and p_2, are very similar whenever the variance of the residuals is very small, but can differ substantially when the parts are only roughly proportional.

Accepting that the two p-values are relevant, an average of them can be an adequate approach. This average is quite peculiar. When exact, p-values are uniformly distributed on $(0,1)$ in a random sampling. An average of them should maintain this property. The way to carry out an average of p-values is using a *probit* transformation. If Φ denotes the standard normal cumulative distribution and p_1, p_2 are two proper p-values the average should be

$$\Phi\left(\frac{1}{2}(\Phi^{-1}(p_1)+\Phi^{-1}(p_2))\right) ,$$

so that the result is still uniformly distributed. In fact, $\Phi^{-1}(p_j)$ has standard normal distribution, and also the average for $j=1,2$; applying Φ again, the distribution is again uniform.

Table 2: p-values of t-student test of proportionality for Moeller et al. 2008 data. Upper triangle: p-value in the regression taking the clr in row as response variable. Lower triangle: p-value in the regression taking the clr in column as response variable.

	Na	K	Mg	Ca	Cl	SO_4	HCO_3
Na		0.0004	0.0001	0.0000	0.0000	0.0000	0.0000
K	0.0000		**0.0266**	0.0000	0.0000	0.0000	0.0000
Mg	0.0000	0.0029		0.0000	0.0000	0.0000	0.0000
Ca	0.0000	0.0012	**0.1286**		0.0000	0.0000	0.0000
Cl	0.0091	0.0078	0.0037	0.0000		0.0000	0.0000
SO_4	**0.0165**	0.0000	0.0000	0.0000	0.0007		0.0000
HCO_3	0.0000	0.0030	0.0073	**0.0262**	0.0000	0.0002	

Table 2 shows the p-values of the t-student test of proportionality. Entries in the upper triangle correspond to regressions taking the clr component of the row element as response variable; alternatively, the clr component of the column element as response variable. Most p-values are less than 0.01, thus indicating rejection of $H_0: \beta_1 = 1$. Asymmetry of p-values is substantial.

Table 3: Φ-averaged p-values and R^2 of t-student test of proportionality for Moeller et al. 2008 data. Upper triangle: R^2. Lower triangle: p-values.

$p\backslash R^2$	Na	K	Mg	Ca	Cl	SO_4	HCO_3
Na	1	0.0218	0.0013	0.0697	0.7810	0.0065	0.6100
K	0.0001	1	0.3048	0.0054	0.0009	0.2147	0.0850
Mg	0.0000	0.0095	1	0.1959	0.0048	0.3277	0.0637
Ca	0.0000	0.0000	0.0031	1	0.1127	0.1639	0.0024
Cl	0.0000	0.0000	0.0000	0.0000	1	0.0311	0.6159
SO_4	0.0000	0.0000	0.0000	0.0000	0.0002	1	0.0557
HCO_3	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1

Table 3 shows that all Φ-averaged p-values are significant, thus suggesting rejection of $H_0: \beta_1 = 1$. The only case for which H_0 could be maintained is the possible, but doubtful, association between Mg and K. It should be noted that R^2 for regression between clr-Cl and clr-Na is relatively high but the slopes of the fitted regression lines seem to be different to 1. The case of clr-Mg and clr-K has a relative low value of R^2, but the highest p-value. These cases are visualized in the scatter-plot shown in Figure 1. In the left panel, two clear outliers (in red) explain the rejection of H_0, as they cause leverage in the regression and fitted lines deviate from unit slope. Right panel shows the sample roughly following two unit slope lines. Regression using the whole sample results in a poor R^2 and deviation from unitary slope of the two regression lines.

The proposed test on unitary slope in the simple regression of two clr-components may be useful for discarding true compositional association across the sample, when the variance of the log-ratio is

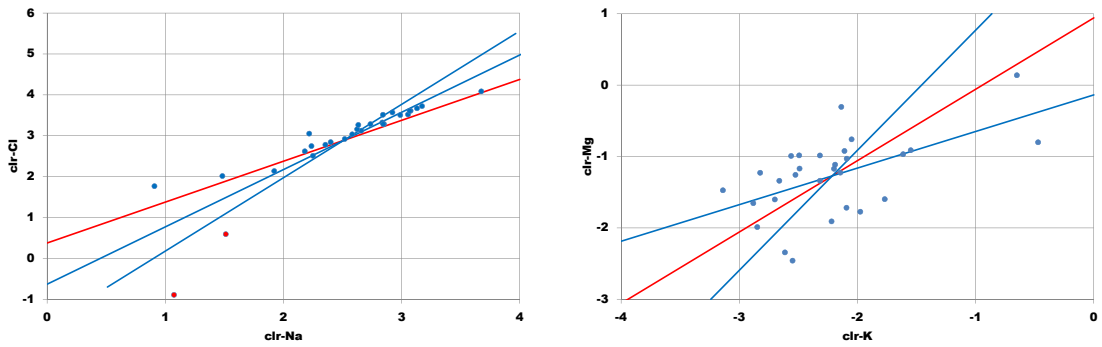


Figure 1: Regression of clr components. Red line: slope equal 1. Blue lines: fitted regression lines. Sample values in blue. Left panel, clr components Na-Cl; two outliers red marker. Right panel, clr components K-Mg.

relatively small. Rejections of H_0 can be due to several reasons, but the presence of high-leverage outliers is one of the most likely reasons.

To address the lack of symmetry in the test regression with techniques that produce invariant estimates under permutation of $\mathrm{clr}_1(\mathbf{x})$ and $\mathrm{clr}_2(\mathbf{x})$ is considered. There are, at least, two alternatives. The first one is the total least squares (or principal component analysis) approach to the model (4). However, it is not appropriate in this case as, in a two part subcomposition, the total variance within the subcomposition and the variance of residuals are equal, whichever the value of β_1, thus invalidating any testing procedure. If the complete composition is considered alternatively, the exact proportionality seldom can be identified with a principal direction. Therefore, total least squares does not seem to be useful in this case. A second alternative is known as regression on the major axis, (Warton et al., 2006), which is explored in a companion paper in these proceedings.

4 Testing regression of log-ratios

As stated in the previous Section 3, the exact compositional association of parts X_1 and X_2 consists in their proportionality $X_1 = a\, X_2$, where a is any positive constant. Taking log's on the proportionality equation yields

$$\log \frac{X_1}{X_2} = \log a \ . \tag{5}$$

Generally, Equation (5) is not fulfilled exactly by the sample values. The question arising is whether the departures from exact association or proportionality are significant or not. To this end, the variability of residuals from Equation (5) should be compared with some alternative source of variability. There are, at least, two possibilities: considering the variability *within* the subcomposition (X_1, X_2) (Section 3), or alternatively, the variability of the *complete* composition. Other sources of variability may be considered, but it requires embedding the problem in a more general context.

If the context of the problem is restricted to the *complete* composition, the composition and their samples can be expressed in ilr-coordinates and, particularly, in balances (Egozcue et al., 2003; Egozcue and Pawlowsky-Glahn, 2005). For the present purposes, the two first balances are taken as

$$b_1 = \frac{1}{\sqrt{2}} \log \frac{X_1}{X_2} \quad , \quad b_2 = \sqrt{\frac{2(D-2)}{D}} \log \frac{\sqrt{X_1 X_2}}{(X_3 \cdot X_4 \cdots X_D)^{1/D-2}} \ ,$$

and the subsequent balances b_j, $j = 3, 4, \ldots, D-1$ are assumed to be orthogonal to the b_1 and b_2 coordinates. The sample balances are denoted b_{ij}, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, D-1$. Inspired on the exact association (5), the regression model

$$b_{i1} = \beta_0 + \sum_{j=2}^{D-1} \beta_j b_{ij} + \varepsilon_i \ , \ D \geq 3 \ , \tag{6}$$

can be compared with (5). In fact, if β_j, $j = 2, 3, \ldots, D-1$ are not significantly different from 0, then Eq. (5) is approximately valid with $\beta_0 = (1/\sqrt{2}) \log a$. Testing proportionality of X_1 and X_2 in this context is equivalent to test the standard regression hypothesis

$$H_0: \ \beta_2 = \beta_3 = \cdots = \beta_{D-1} = 0 \quad , \quad H_1: \ \text{otherwise.}$$

Under the reasonable assumption of normality of residuals, $\varepsilon_i \sim N_{\text{ind}}(0, \sigma_\varepsilon^2)$, the standard F-test can be used for testing H_0. The F-test statistics is

$$F = \frac{\frac{1}{n-D+1} \sum_i (b_{i1} - \widehat{b_1})^2}{\frac{1}{n-1} \sum_i (b_{i1} - \bar{b_1})^2} = \frac{\frac{1}{n-D+1} \widehat{\text{Var}}[\varepsilon]}{\frac{1}{n-1} \widehat{\text{Var}}[b_1]} \ ,$$

where $\bar{b_1}$ is the sample average of the response b_{i1} and $\widehat{b_1}$ is the estimated linear predictor of b_1, i.e. $\widehat{b_1} = \widehat{\beta_0} + \sum_{j=2}^{D-1} \widehat{\beta_j} b_{ij}$. Under normality of residuals, the test statistic F is distributed $F_{n-D-1,n-1}$ which allows the computation of a p-value for H_0 and consequently for association of the parts X_1 and X_2.

The proposed F-test checks correlation between the residuals under H_0 with the other balances b_2, b_3, ... An almost constant b_1 is expected to be uncorrelated to any other source of variation, thus matching H_0. However, such an *almost constant* b_1 can be correlated to other *almost constant* combinations of other balances. In these cases the rejection of H_0 points out that the association between parts can be extended to larger groups of parts. Therefore, rejection of H_0 can be due to several reasons, the main of which is that association is not *only of two parts* but possibly of larger groups of parts. This test is called log-ratio regression test (lrRT).

When lrRT's are run on the Moeller et al. data, all paired regressions are significative, thus rejecting strong association between any couple of parts. The least significant couples are: Ca-Mg (p-value 0.024), Ca-SO_4 (p-value 0.004) and Cl-Na (p-value 0.001). This means that departures from proportionality are partially predictable from other balances. In the case of Cl-Na (Fig. 1, left panel), this means that the two outliers filled in red are partially predictable from balances different from that of Cl-Na, thus suggesting a larger association than just Cl-Na.

5 Exploring compositional association of two parts

The previous example, using major anions and cations from Moeller et al. (2008), led to doubtful associations of some pairs of elements. The significant results of the clrST and lrRT discarded any strong association. However, the possible and intuitive association of Na and Cl was disturbed by two outliers. After removing these two outliers, the remaining data set (27 compositions) is now used to review a case in which two clear compositional associations are found out.

One of the first steps in exploratory analysis of compositional data starts examining the variation matrix. Table 4 shows $\tilde{\mathbf{V}}$, the normalized version of the variation matrix in the lower triangle and the original variation matrix in the upper triangle. A first observation is that the original data set had total variance 7.583; after the removal of the two outliers total variance is remarkably reduced to 5.762.

In order to make the table readable, normalized variance values larger than 1 have been printed in red. Any compositional association for the corresponding couples of parts should be discarded. Normalized variances less than 1 suggest some kind of association, but only values less than 0.3 are presented in blue, thus remarking the possibility of a meaningful association. Values under 0.2 have been typed in boldface, as corresponding parts are serious candidates of association. For normalised variance values less than 0.3 (in blue), the p-values of clrST are marked with the symbol $*$ and of the lrRT with the symbol †. A significant p-value tends to reject association, consequently p-values are coded as: 3 symbols if $p > 0.1$; 2 symbols if $0.1 \geq p > 0.05$; 1 symbol if $0.05 \geq p > 0.01$. An inspection of Table 4 reveals that only two pairs of elements are serious candidates to compositional association, i.e. K-Mg (normalised variance 0.160) and Na-Cl, with a much less normalised variance 0.010. The next step should be checking whether the two tests reject association. In the case K-Mg, the p-value of clrST is 0.29, corresponding to $***$, and an almost null p-value in the lrRT. The

Table 4: Variation matrix of Moeller et al. 2008 data after removing two outliers. Upper triangle is the original variation matrix; the lower triangle is the normalised version. Normalized variances less than 0.30 are presented in blue. Normalized variances less than 0.20 are presented in bold face. Normalized variances greater than 1 in red. For normalized variances less than 0.30, non-significance of proportionality ($*$) and regression ($\dagger$) tests is marked with: 3 symbols if $p > 0.1$; 2 symbols if $0.1 \geq p > 0.05$; 1 symbol if $0.05 \geq p > 0.01$. Total variance is 5.762.

$\tilde{\mathbf{V}} \setminus \mathbf{V}$	Na	K	Mg	Ca	Cl	SO_4	HCO_3
Na	0	0.478	0.560	0.847	0.019	1.638	4.424
K	0.249 **	0	0.307	0.844	0.445	2.080	4.112
Mg	0.292 *	**0.160** ***	0	0.498	0.469	2.140	3.947
Ca	0.441	0.439	0.259 ***††	0	0.728	2.188	4.176
Cl	**0.010** ***†	0.232 ***	0.244 **	0.379	0	1.723	4.510
SO_4	0.853	1.083	1.114	1.139	0.897	0	4.200
HCO_3	2.303	2.141	2.055	2.174	2.348	2.187	0

conclusion would be *there is a noisy association between K and Mg, but probably it can be included in a more complex association including more elements*. The case Na-Cl is a clear association due to its very low normalised variance, confirmed by a p-value of 0.94 ($***$) in the clrST. However, the lrRT gives a p-value 0.02, i.e. there is some kind of linear relationship of the residuals with other balances, thus suggesting a slight compositional association with other elements. The conclusion would be *there is a strong compositional association between Na and Cl, but there is an indication that departures from proportionality are related with other elements in the composition.*

A further visual inspection of Table 4 reveals another possibility of compositional association, namely the pair Mg-Ca. The normalised variance is larger than in the two previous cases, but the two tests (clrST, lrRT) are almost non-significant (0.158 $***$, 0.098$\dagger\dagger$). The squared correlation coefficient in the clrST is $R^2 = 0.16$, that reveals a poor association, although the unit slope of clr-Mg / clr-Ca cannot be rejected. The conclusion is that the two tests are not useful to establish compositional association, but useful to reject it.

6 Conclusion

Compositional association between two parts of a composition has been defined as proportionality of the two parts. Compositional association has been described using the variation matrix from the beginning of compositional data analysis. However, scaling of the variation matrix for exploratory analysis is still a pending question. Such a normalization has been proposed.

Also, statistical checking of compositional association has been explored. Two tests are proposed, the clr-slope test (clrST) is based on the fact that exact compositional association implies that the two clr components must lay over a unit slope line. Under normality of residuals, a `t`-statistics is then used. Significant values of the statistics suggest rejecting the compositional association. This test is sensible to levering outliers able to deviate the slope of the line.

A second test, log-ratio regression test (lrRT), is based in the idea that, if a log-ratio of two elements is constant, then it is uncorrelated with any other balance orthogonal to the log-ratio. The regression model uses the log-ratio of the two parts as response variable and the rest of orthonormal coordinates as explanatory. The test is performed, under normality of the residuals, using the standard F-test in regression analysis. A significant p-value is interpreted as a suggestion of compositional association of more than two parts. It should be remarked that non-significant results in the two tests do not support compositional association. They are only useful to discard associations previously suggested by a small variance of the corresponding log-ratio.

The presentation of the results in a table is also proposed. The table shows the variation matrix

and its normalised version, together with symbols indicating the non-significance of the two tests proposed.

Acknowledgements

This research has been supported by the Spanish Ministry of Education, Culture and Sports under a Salvador de Madariaga grant (Ref. PR2011-0290); by the Spanish Ministry of Economy and Competitiveness 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 project Ref: 2009SGR424.

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.

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 (2011). Análisis composicional de datos en ciencias geoambientales. *Boletín Geológico y Minero 122*(4), 439–452.

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.

Lovell, D., V. Pawlowsky-Glahn, and J. J. Egozcue (2013, March). Don't correlate proportions! *http://www.slideshare.net/AustralianBioinformatics/dont-correlate-proportions.*.

Mateu-Figueras, G., V. Pawlowsky-Glahn, and J. J. Egozcue (2011). The principle of working on coordinates. In V. Pawlowsky-Glahn and B. A. (Eds.), *Compositional Data Analysis: Theory and Applications*, pp. 31–42. Wiley, Chichester UK.

Moeller, P., S. M. Weise, M. Tesmer, P. Dulsky, A. Pekdegger, U. Bayer, and F. Magri (2008). Salinization of groundwater in the North German Basin: results from conjoint investigation of major, trace element and multi-isotope distribution. *Int. J. Earth Sciences 97*, 1057–1073.

Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2011). *Principal Balances*. In: Egozcue, J.J., Tolosana-Delgado, R. and Ortego, M.I. (eds.), *Proceedings of CoDaWork'2011, The* 4^{th} *International Compositional Data Analysis Workshop*, CIMNE, Barcelona (E).

Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2013). Principal balances. *Mathematical Geosciences*, (submitted).

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.

Warton, D. I., I. J. Wright, D. S. Falster, and M. Westoby (2006). Bivariate line-fitting methods for allometry. *Biological Reviews 81*(2), 259–291.

Statistical analysis of compositional 2×2 tables

K. FAČEVICOVÁ[12], K. HRON[12]

[1]Department of Mathematical Analysis and Applications of Mathematics - Palacký University, Czech Republic
kamila.facevicova@gmail.com
[2]Department of Geoinformatics - Palacký University, Czech Republic

1 Introduction

A 2×2 compositional table could be considered as a special case of four-part composition, which represents relationship between two (row and column) factors. It is an continuous analogy to the well-known contingency tables. Important is to point out that cells of compositional tables express quantitatively contributions on the total and thus only ratios between parts are important source of information in the table.

As well as in the case of contigency tables, the main question is whether there exists any relationship between factors or not. This relationship could be analyzed through decomposition of the original table onto its independent and interaction parts. If row and column factors are independent, the whole information about the original table is contained in the independent part. The corresponding hypothesis could be tested using a sample of compositional tables and proper choice of coordinates. Since 2×2 compositional tables analysis is the low-dimensional case of general $I \times J$ compositional tables analysis it is also possible to visualize the real data structure in orthonormal coordinates.

2 Compositional tables and their geometrical properties

Since compositional tables are special case of D-part compositions, as will be shown later, the main properties of compositional tables could be presented on this general case. Recall that a random D-part composition is a row vector $\mathbf{x} = (x_1, x_2, \ldots, x_D)$, where $x_i > 0, i = 1, \ldots, D$, and each part quantitatively describes its contribution on the whole (Egozcue, 2009). Definition of compositional data enables their scaling to a prescribed constant sum constraint κ (i.e. to 1 in case of proportions and 100 for percentages) without loss of information; formally, we refer to a closure operation and denote

$$\mathcal{C}(\mathbf{x}) = \left(\frac{\kappa \cdot x_1}{\sum_{i=1}^{D} x_i}, \frac{\kappa \cdot x_2}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa \cdot x_D}{\sum_{i=1}^{D} x_i} \right).$$

Thus, the sample space of representations of compositional data is the simplex, a $(D-1)$-dimensional subset of $\mathbf{R}^D$ defined as

$$S^D = \left\{ \mathbf{x} = (x_1, x_2, \ldots, x_D) |\ x_i > 0,\ i = 1, 2, \ldots, D;\ \sum_{i=1}^{D} x_i = \kappa \right\}.$$

Any analysis applied to compositional data should fulfill three main conditions (Egozcue, 2009):

- Relative scale: Ratios express the differences between observations rather than absolute distances.
- Invariance: Scaling of the original data should not alter the results of the analysis.
- Subcompositional coherence: Results obtained from a composition of D parts should not be in contradiction with results that are obtained from a sub-composition containing d parts, $d < D$.

Specifically, a four-part composition $\mathbf{x} = \mathcal{C}(x_{11}, x_{12}, x_{21}, x_{22})$ can be re-ordered into a 2×2 compositional table

$$\mathbf{x} = \mathcal{C} \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{pmatrix},$$

which represents the relationship between the row and column factors.

Basic operations with compositional data follow the Aitchison geometry. Perturbation and power transformation for 2×2 compositional tables $\mathbf{x}$ and $\mathbf{y}$ and a real number α, respectively, are thus defined as

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C} \begin{pmatrix} x_{11}y_{11} & x_{12}y_{12} \\ x_{21}y_{21} & x_{22}y_{22} \end{pmatrix}, \quad \alpha \odot \mathbf{x} = \mathcal{C} \begin{pmatrix} x_{11}^{\alpha} & x_{12}^{\alpha} \\ x_{21}^{\alpha} & x_{22}^{\alpha} \end{pmatrix} \quad .$$

Note that $\mathbf{n} = \mathcal{C} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ is the neutral element in the Abelian group $(\mathcal{S}^D, \oplus)$. The Aitchison inner product of two factor compositional tables $\mathbf{x}$ and $\mathbf{y}$ is defined as $\langle \mathbf{x}, \mathbf{y} \rangle_a =$

$$\frac{1}{4}\left(\ln\frac{x_{11}}{x_{12}}\ln\frac{y_{11}}{y_{12}} + \ln\frac{x_{11}}{x_{21}}\ln\frac{y_{11}}{y_{21}} + \ln\frac{x_{11}}{x_{22}}\ln\frac{y_{11}}{y_{22}} + \ln\frac{x_{12}}{x_{21}}\ln\frac{y_{12}}{y_{21}} + \ln\frac{x_{12}}{x_{22}}\ln\frac{y_{12}}{y_{22}} + \ln\frac{x_{21}}{x_{22}}\ln\frac{y_{21}}{y_{22}}\right) \quad .$$

Denote $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus [(-1) \odot \mathbf{y}]$ as the inverse operation of perturbation. Then from the Euclidean vector space properties of the Aitchison geometry,

$$\|\mathbf{x}\|_a = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a} \quad \text{a} \quad d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a \quad ,$$

represents the Aitchison norm of a table $\mathbf{x}$ and the distance between two tables $\mathbf{x}$ and $\mathbf{y}$, respectively.

As was mentioned above, sample space of 2×2 compositional tables is simplex $\mathcal{S}^4$, the 3-dimensional subset of $\mathbf{R}^4$ and according to Egozcue and Pawlowsky-Glahn (2005), the original table $\mathbf{x}$ could be projected onto arbitrary subspace using formula $\langle \mathbf{x}, \mathbf{E} \rangle_a \odot \mathbf{E}$, where $\mathbf{E}$ forms an orthonormal basis in this subspace. First important group constitute of projections onto row subspaces $S^4(\mathrm{row}_1)$ and $S^4(\mathrm{row}_2)$,

$$\mathrm{row}_1(\mathbf{x}) = \mathcal{C} \begin{pmatrix} x_{11} & x_{12} \\ \sqrt{x_{11}x_{12}} & \sqrt{x_{11}x_{12}} \end{pmatrix} \quad \text{and} \quad \mathrm{row}_2(\mathbf{x}) = \mathcal{C} \begin{pmatrix} \sqrt{x_{21}x_{22}} & \sqrt{x_{21}x_{22}} \\ x_{21} & x_{22} \end{pmatrix}.$$

These projections extract from the original table information about first and second row, respectively. Analogously, projections onto $S^4(\mathrm{col}_1)$ and $S^4(\mathrm{col}_2)$

$$\mathrm{col}_1(\mathbf{x}) = \mathcal{C} \begin{pmatrix} x_{11} & \sqrt{x_{11}x_{21}} \\ x_{21} & \sqrt{x_{11}x_{21}} \end{pmatrix}, \quad \mathrm{col}_2(\mathbf{x}) = \mathcal{C} \begin{pmatrix} \sqrt{x_{12}x_{22}} & x_{12} \\ \sqrt{x_{12}x_{22}} & x_{22} \end{pmatrix}.$$

extract information about columns.

For compositional tables analysis also projections onto subspaces $\mathcal{S}^4(\mathrm{row}^{\perp})$ and $\mathcal{S}^4(\mathrm{col}^{\perp})$ which are orthogonal to both row and column subspaces, respectively are important. These projections have the form

$$\mathrm{row}^{\perp}(\mathbf{x}) = \mathcal{C} \begin{pmatrix} \sqrt{x_{11}x_{12}} & \sqrt{x_{11}x_{12}} \\ \sqrt{x_{21}x_{22}} & \sqrt{x_{21}x_{22}} \end{pmatrix} \quad \text{and} \quad \mathrm{col}^{\perp}(\mathbf{x}) = \mathcal{C} \begin{pmatrix} \sqrt{x_{11}x_{21}} & \sqrt{x_{12}x_{22}} \\ \sqrt{x_{11}x_{21}} & \sqrt{x_{12}x_{22}} \end{pmatrix}$$

and preserve only information about relationship between rows or columns, respectively (Fišerová and Hron, 2011). These two projections are mutually orthogonal and since all presented subspaces have dimension 1, the original table could be decomposed by

$$\mathbf{x} = \mathrm{row}^{\perp}(\mathbf{x}) \oplus (\mathrm{row}_1(\mathbf{x}) \oplus \mathrm{row}_2(\mathbf{x})) \quad \text{or} \quad \mathbf{x} = \mathrm{col}^{\perp}(\mathbf{x}) \oplus (\mathrm{col}_1(\mathbf{x}) \oplus \mathrm{col}_2(\mathbf{x})) \,.$$

Information contained in $\mathrm{row}^{\perp}$ and $\mathrm{col}^{\perp}$ are also sufficient to reconstruct compositional table when there is no relationship between factors. Subspace of these tables, $\mathcal{S}^4_{ind}$, arises from a perturbation of subspaces $\mathcal{S}^4(\mathrm{row}^{\perp})$ and $\mathcal{S}^4(\mathrm{col}^{\perp})$ and has dimesion 2. The independence table is then projection of the original table $\mathbf{x}$ onto this subspace,

$$\mathbf{x}_{ind} = \mathrm{row}^{\perp}(\mathbf{x}) \oplus \mathrm{col}^{\perp}(\mathbf{x}) = \mathcal{C} \begin{pmatrix} x_{11}\sqrt{x_{12}x_{21}} & x_{12}\sqrt{x_{11}x_{22}} \\ x_{21}\sqrt{x_{11}x_{22}} & x_{22}\sqrt{x_{12}x_{21}} \end{pmatrix} \quad .$$

The remaining information about $\mathbf{x}$, especially about interactions, are included in so-called interaction table $\mathbf{x}_{int}$ and the original table could be then decomposed as

$$\mathbf{x} = \mathbf{x}_{ind} \oplus \mathbf{x}_{int} \quad .$$

From this decomposition we get interaction table in form

$$\mathbf{x}_{int} = \mathbf{x} \ominus \mathbf{x}_{ind} = \mathcal{C}\begin{pmatrix} \frac{1}{\sqrt{x_{12}x_{21}}} & \frac{1}{\sqrt{x_{11}x_{22}}} \\ \frac{1}{\sqrt{x_{11}x_{22}}} & \frac{1}{\sqrt{x_{12}x_{21}}} \end{pmatrix} = \mathcal{C}\begin{pmatrix} \sqrt{x_{11}x_{22}} & \sqrt{x_{12}x_{21}} \\ \sqrt{x_{12}x_{21}} & \sqrt{x_{11}x_{22}} \end{pmatrix} \quad .$$

This decomposition of compositional table onto its independent and interaction parts enables to perform statistical inference on relationship between factors. This inference is much simplified when a proper choice of coordinates is used as it will be shown in the next section.

3 Statistical analysis of compositional tables in coordinates

The main challenge relating to 2×2 compositional tables is to test whether row and column factors are independent or not. The null hypothesis that factors are independent can be reformulated using decomposition of the original table, i.e. all information about the compositional table $\mathbf{x}$ is concentrated in the independent table $\mathbf{x}_{ind}$ and the centre of the distribution of $\mathbf{x}_{int}$ equals to neutral element $\mathbf{n}$. If we assume normal distribution of the (random) compositional table $\mathbf{x}$ on the simplex (Mateu-Figueras and Pawlowsky-Glahn, 2008), we can test this hypothesis against alternative that centre of $\mathbf{x}_{int}$ is not the neutral element. Since normal distribution on the simplex presumes normality in orthonormal coordinates with respect to the Aitchison geometry it seems to be easier to perform the test in the orthonormal coordinates (Egozcue et al., 2003; Egozcue and Pawlowsky-Glahn, 2005) rather than with the original compositional tables.

Orthonormal coordinates are assigned to a composition $\mathbf{x} = \mathcal{C}(x_1, \ldots, x_D) \in \mathcal{S}^D$ through isometric (ilr) logratio transformation which results in $(D-1)$-dimensional real vector $\mathbf{z} = h(\mathbf{x}) = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_a, \ldots, \langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a) = (z_1, z_2, \ldots, z_{D-1})$, where $\mathbf{e}_i = \mathcal{C}(e_{i1}, \ldots, e_{i,D})$, $i = 1, \ldots, D-1$ form an orthonormal basis on the simplex. It is easy to see that the real vector $\mathbf{z}$ represents orthonormal coordinates in the real space $\mathbb{R}^{D-1}$. Consequently, the ilr transformation is an isometric isomorphism, i.e. for $\mathbf{x}_1, \mathbf{x}_2, \in S^D, \alpha, \beta \in \mathbb{R}$ the following properties hold

$$h(\alpha \odot \mathbf{x}_1 \oplus \beta \odot \mathbf{x}_2) = \alpha \cdot \mathbf{z}_1 + \beta \cdot \mathbf{z}_2, \quad \langle \mathbf{x}_1, \mathbf{x}_2 \rangle_a = \langle \mathbf{z}_1, \mathbf{z}_2 \rangle.$$

The concrete orthonormal (ilr) coordinates $\mathbf{z}$ depend on a particular orthonormal basis of $\mathcal{S}^D$ which was chosen for their construction. Usually their easy interpretability is desired; construction using sequential binary partition (SBP) represents one possibility of satisfying this requirement (Egozcue and Pawlowsky-Glahn, 2005). Although different partitions lead to different coordinates these coordinates are just an orthogonal transformation of each other (Egozcue et al., 2003) and the result of analysis does not depent on the concrete choice of the SBP.

SBP	x_{11}	x_{12}	x_{21}	x_{22}	r	s
Step 1	+	−	−	+	2	2
Step 2		+	−		1	1
Step 3	+			−	1	1

Table 1: Tabular representation of SBP.

For testing of independence in 2×2 compositional tables the best choice of coordinates results from partition which separate the diagonal parts from the rest in the first step as it is shown in Table 1. This partition leads to two nonzero coordinates for the independent table and only one nonzero coordinate for the interaction table what corresponds to dimensions of subspaces of these tables. All coordinates are listed in Table 2. From the isometry of the ilr transformation it follows that the decomposition of $\mathbf{x}$ into independent and interaction tables, $\mathbf{x} = \mathbf{x}_{ind} \oplus \mathbf{x}_{int}$, can be expressed in coordinates as $\mathbf{z} = \mathbf{z}_{ind} + \mathbf{z}_{int}$ and the null hypothesis on independence could be reformulated again, i.e. the mean value of $\mathbf{z}_{int}$ is a null vector. And since there is only one nonzero coordinate for $\mathbf{z}_{int}$ the hypothesis on independence between two factors in compositional table can be tested using the well-known t-test.

$\mathbf{z}$	$\mathbf{z}_{ind}$	$\mathbf{z}_{int}$
$\frac{1}{2}\ln\frac{x_{11}x_{22}}{x_{12}x_{21}}$	0	$\frac{1}{2}\ln\frac{x_{11}x_{22}}{x_{12}x_{21}}$
$\frac{\sqrt{2}}{2}\ln\frac{x_{12}}{x_{21}}$	$\frac{\sqrt{2}}{2}\ln\frac{x_{12}}{x_{21}}$	0
$\frac{\sqrt{2}}{2}\ln\frac{x_{11}}{x_{22}}$	$\frac{\sqrt{2}}{2}\ln\frac{x_{11}}{x_{22}}$	0

Table 2: Ilr coordinates of $\mathbf{x}$, $\mathbf{x}_{ind}$ and $\mathbf{x}_{int}$ using SBP.

4 Example - The theft detection

Methodology of compositional tables analysis could be presented on example of the theft detection. For this purpose data from 73 cities in Czech Republic are available, which describe relationship between the type of theft and its detection in the same year as the theft was committed. We are observing two factors. The first of them is type of theft: if there were an obstacle that had to be overcome we are talking about burglary, otherwise it is so-called simple theft. The second factor devotes whether the case was explained or not. An example of table we are dealing with is displayed in Table 3.

Prague	Explained	Not explained
Simple theft	158	2992
Burglary	29	899

Prague	Explained	Not explained
Simple theft	0.039	0.734
Burglary	0.007	0.220

Table 3: Distribution of the theft detection in Prague in 2011 in absolute numbers (left) and in proportions (right).

Following the methodology presented above all tables were decomposed onto their independent and interaction parts. In the case of Prague the new tables were

$$\mathbf{x}_{ind}^{Prague} = \begin{pmatrix} 0.032 & 0.778 \\ 0.008 & 0.183 \end{pmatrix} \quad \mathbf{x}_{int}^{Prague} = \begin{pmatrix} 0.281 & 0.219 \\ 0.219 & 0.281 \end{pmatrix}.$$

Note that these tables follow the condition $\mathbf{x}_{ind}^{Prague} \oplus \mathbf{x}_{int}^{Prague} = \mathbf{x}^{Prague}$. All sets of tables (original, independent and interaction) were expressed in orthonormal coordinates in the case of Prague to

$$\mathbf{z}^{Prague} = (0.246, 3.278, -1.229),$$

$$\mathbf{z}_{ind}^{Prague} = (0, 3.278, -1.229) \qquad \text{and} \qquad \mathbf{z}_{int}^{Prague} = (0.246, 0, 0).$$

It is obvious that the relation $\mathbf{z}^{Prague} = \mathbf{z}_{ind}^{Prague} + \mathbf{z}_{int}^{Prague}$ holds again. Finally, the sample that consists of 73 first coordinates of $\mathbf{z}_{int}$ was tested for null mean value. This hypothesis was rejected by the t-test (p-value $= 7.627 \times 10^{-5}$) so it could be also rejected that the detection does not depend on the type of theft. Analogous results could be obtained by testing smaller data set consisting of 14 tables which summarize data from cities by the region to which they belong. The hypothesis on independence was rejected again, only the p-value has increased to 0.003. Because of the ilr coordinates corresponding to SBP form a representation of a compositional table in real space it is convenient in the case of regions to display the first coordinates of the interaction tables and the last two coordinates of the independent tables in Figure 1. The dotplot of the coordinates of the interaction tables (left) shows that the lowest contribution to the hypothesis rejection has the Central Bohemian Region whose coordinate is nearest to the origin. On the other hand, the highest contribution has the Hradec Kralove Region which has the highest value of the coordinate (0.43) and is situated on the right side of the plot, farthest from the origin. The coordinate of the Liberec Region has negative value quite far from the origin (-0.171). Since the only nonzero coordinate of interaction table is defined as the logarithm of the ratio between the chance to explanation of the simple theft and the chance to explanation of the burglary, divided by 2, the odds ratio (Agresti, 2002) in the Liberec Region is $e^{2\cdot(-0.171)} = 0.71$ and so there is a higher chance to explain the burglary then the simple theft. The situation is a reverse one in regions whose interaction table coordinate has positive value like in case of the Hradec Kralove Region where the

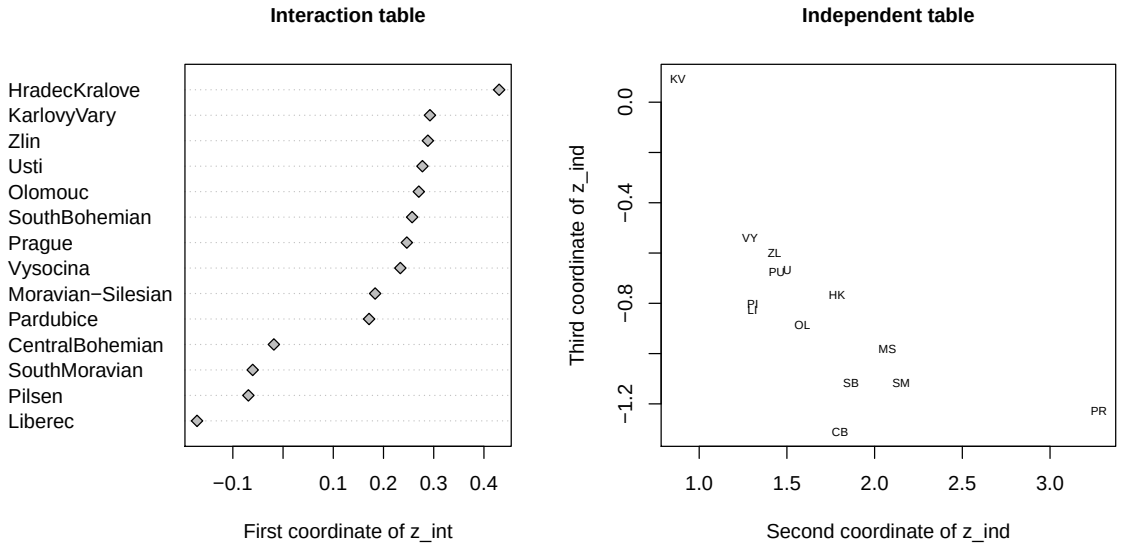


Figure 1: Comparison of regions by ilr coordinates of interaction (left) and independent table (right).

odds ratio equals to 2.364 and so the chance to explain the simple theft is more than twice higher than the chance to explanation of the burglary. The right figure shows coordinates of the independent table. Regions with the lowest values of these coordinates are concentrated in the upper left corner of the figure, the most information about these regions is then conveyed by the interaction table. All regions have positive second coordinate; by its construction this means that there are more unexplained simple thefts then explained burglaries and this ratio grows, the more is the region situated on the right side of the figure. The most of the regions also have negative third coordinate and thus there is more not explained burglaries then explained simple thefts and this ratio also grows, the more is the region situated on the bottom of the graph. Two extreme cases are the Karlovy Vary Region and Prague. There occure a bit more of unexplained simple thefts than explained burglaries in Karlovy Vary. There is also the same amount of unexplained burglaries as simple thefts. Nevertheless, there are much more not explained simple thefts in Prague than explained burglaries and also more unexplained burglaries than explained simple thefts. We can conclude that the police in here is quite unsuccessful as for explaining thefts, regardless of their type and so there is the strongest independence between the type of theft and its detection which also corresponds to the location of Prague in both figures.

Acknowledgements Authors gratefully acknowledge the support of the Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic).

References

Agresti, A. (2002). *Categorical data analysis (2 ed.).* J. Wiley & Sons, New York.

Aitchison, J. (1986). *The statistical analysis of compositional data.* Chapman and Hall, London.

Aitchison, J., C. Barceló-Vidal, J.A. Martín-Fernández, V. Pawlowsky-Glahn (2000). *Logratio analysis and compositional distance.* Math Geol 32:271-275.

Egozcue, J. J. (2009). *Reply to "On the Harker Variation Diagrams" by J.A. Cortés.* Math Geosci 41:829-834.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, C. Barceló-Vidal (2003). *Isometric logratio transformations for compositional data analysis.* Math Geol 35:279–300.

Egozcue, J. J., V. Pawlowsky-Glahn (2005). *Groups of parts and their balances in compositional data analysis.* Math Geol 37:795–828.

Fišerová, E., K. Hron (2011). *On interpretation of orthonormal coordinates for compositional data.* Math Geosci 43:455–468.

Mateu-Figueras, G., V. Pawlowsky-Glahn (2008). *A critical approach to probability laws in geochemistry.* Math Geosci 40:489–502.

Effect of processing on the Manzanilla and Hojiblanca green Spanish-style table olive fat as assessed by compositional data analysis

A. LÓPEZ-LÓPEZ[1], A. CORTÉS-DELGADO[2] and A. GARRIDO-FERNÁNDEZ[1*]

[1]Departamento de Biotecnología de Alimentos – Instituto de la Grasa (CSIC), Spain
[2]Sevicio de Analisis al exterior – Instituto de la Grasa (CSIC), Spain
*garfer@cica.es

There are only a few studies related to the changes in the fat composition during table olive processing. A first approach, using conventional multivariate analysis, showed that the oil from ripe olives could suffer diverse transformations during processing, mainly during the storage phase. Usually, fatty acid composition is expressed as percentages, constituting a clear case of compositional data (Aitchison, 1986). However, the effect of treatments, diverse origins, and olive class are still studied by traditional multivariate techniques. In this work, the application of compositional analysis is applied for the first time to investigate the effect of processing on the fatty acid composition of the Manzanilla and Hojiblanca cultivars processed as green Spanish-style table olives. Exploratory data analysis by biplot or balance-dendogram (Pawlowsky-Glahn et al., 2011), discriminat analysis (Templ et al., 2011), and clasical multivariate techniques (applied to raw data and ilr coordinates) were used to disclose changes due to processing or differences between cultivars.

1 Introduction

Table olives are a typical food in the countries around the Mediterranean Basin. However, their consumption has also extended to America, South Africa, Australia and Japan. Nowadays, it is the largest fermented vegetable product in western countries. Its world production reached 2,210,000 tons in the 2009/2010 season (IOC, 2011). Approximately 60% of table olives are processed as green Spanish-style.

The composition of table olives has been studied during the last decades of the last century and the importance of fat in the olive flesh is well established (Fernández Díez et al., 1985; Garrido Fernández et al., 1997).

A detailed study of the fat composition of the Spanish cultivars, according to elaboration styles was carried out by López López et al. (2006); the major fatty acids were C18:1c, C16:0, C18:2n-6 and C18:0. The fatty acid compositions found were similar to those found in olive oil from the respective cultivars (Aparicio & Harwood, 2003). The proportions of saturated, monounsaturated, polyunsaturated and *trans* fat were, on average, 2.07-2.99g/100g flesh, 5.67-19.42g/100flesh, 0.52-3.87g/100g, and 0.08-0.44g/100g flesh, respectively (López López et al., 2006). It has been traditionally thought that the fat was well protected in the interior of the olive cells and was not affected by the different processing operations. However, this assumption has not been confirmed with experimental data. The first studies on fat transformations during ripe olive processing, the most likely fat to be affected by the numerous alkali treatments, washings and thermal sterilization applied to the fruits (Garrido Fernández et al., 1997) showed that acidity, peroxide value, K_{270} and ΔK increased during storage.Most of the fatty acids, triacylglygerols, diacylglycerols, and monoacylglycerols also had significant changes during processing (López López et al., 2009), mainly during the previous storage phase (López López et al., 2010). Traditionally, the fatty acid or triacylglycerides have been used to study the effects of diverse factors (origin, storage, etc) on olive oil composition. Triacylglycerides were used by Vlahov (1996) to classify olive oils according to geographical origin. A statistical multi-test on the triacylglyceride fraction was used as a confirmatory test to distinguish between imported and

exported oils in Italy (Gambacorta et al., 2002). Chemometric techniques (principal component analysis, PCA, and discriminant analysis, DA) were used by Lee et al. (1998) for the characterization of fatty acid composition in vegetable oils. Usually, the studies have applied conventional multivariate techniques to fatty acid, triglycerides or sterol contents, expressed as percentages.

The aim of this work was to compare the results obtained in the statistical analysis of the fatty acid composition of the fats from the Manzanilla and Hojiblanca cultivars subjected to green Spanish-style processing. To simplify the study, only the major fatty acids have been considered in this work.

2 Materials and Methods

2.1 Olives and processsing

Olives were of the Manzanilla and Hojiblanca cultivars harvested at the so-called green maturation stage (JOLCA S.L., Huevar, Sevilla,Spain). They were processed according to the Spanish-style. Briefly, the process consisted of treating the olives with a 1.8 g/100mL NaOH solution until the alkali reached 2/3 of the flesh, followed by one washing with tap water for 18h. Then, the olives were brined in a 9g/100mL NaCl solution. The process followed a typical spontaneous fermentation. After three months, the NaCl concentration was increased to up to 5.5g/100mL NaCl. At the end of the storage period, the olives were packed in glass containers (0.5 g/100mL lactic acid and 5.5 g/100mL NaCl, at equilibrium). The processing and packing may be considered to have followed the standard methodology applied to elaborate this cultivar at industrial scale.

2. 2 Fat extraction

Oil was extracted from the olives according to the ABENCOR system. The olives were pitted, mixed with a homogenizer Ultraturax T25 (IKA-Labortecnik, Staufen, Deutschland) and then boiling water (100 ºC) was added to the paste. The resulting suspension was subjected to malaxation for 40 min at room temperature (22 ºC±2) and the liquid was removed by centrifugation using ABENCOR equipment (Abengoa, Madrid, Spain), similar to that used for the estimation of olive oil yield in olive mills (Martínez et al, 1975). The liquid phase was allowed to decant and the oil was obtained, filtered and subjected to analysis. This method was used to prevent changes in the oil quality as much as possible.

2.3 Fatty acid analysis

The determination of fatty acids was accomplished through the quantification of their methyl esters (FAMES) by GC. The methylation of the fat extracts was performed by heating the fat (100 mg) with 4 mL of 0.2 N sodium methylate in methanol, followed by heating in an acidic medium (Commission Regulation (EEC) 2568/91). The fatty acid methyl esters were then analyzed with a Hewlett-Packard 5890 Series II gas chromatograph, using a fused silica capillary column Select FAME (100mx0.25mmx0.25μm film thickness) and a flame ionization detector. The fatty acid results were expressed as the percentage of area of the respective chromatograms.

2.4 Statistical analysis

Data were analyzed using Statistica (release 6.0) (StatSoft, Inc., Tulsa, OK, USA) for the classical multivariate analysis, CodaPack (Comas-Cufí and Thió-Henestrosa, 2011) and robCompostions (Templ et al., 2011) in the case of the application of compositional data analysis.

3 Results

The results of the analysis of the fat extracted from Manzanilla and Hojiblanca after each processing step are shown in Table l. As expected, the sum of the fatty acids was fairly close to 100 because only those fatty acids detected in low proportions were removed from the original data base. These data are a typical compositional data base. However, as commented above, similar information has been widely used to estimate the effect of diverse variables, region or country of origin,etc. on olive oil composition The data were subjected to analysis to estimate the summary statistics.

Table 1. Fatty acid composition of olive fat (Manzanilla and Hojiblanca) after each processing step (cult., cultivar; treat., treatment). Fatty acids in low proportions were removed from the original data base.

Cultivar	Cult vs treat.	Treat.	C 16:0	C 17:0	C 18:0	C 20:0	C 21:0	C 16:1	C 17:1	C 18:1c	C 18:2n-6	C 18:3n-3	Others	Total
Manzanilla	MT0	T0	14.259	0.159	2.851	0.390	0.011	1.520	0.278	65.672	6.396	0.811	.....	92.348
Manzanilla	MT0	T0	14.167	0.156	2.843	0.399	0.011	1.499	0.270	65.703	6.380	0.768	.....	92.195
Manzanilla	MT1	T1	14.267	0.159	2.888	0.404	0.011	1.468	0.267	66.841	5.706	0.703	.....	92.715
Manzanilla	MT1	T1	13.968	0.161	2.892	0.410	0.013	1.485	0.273	67.298	5.781	0.718	.....	92.999
Manzanilla	MT1	T1	14.793	0.158	2.808	0.396	0.013	1.501	0.268	66.886	5.893	0.746	.....	93.462
Manzanilla	MT1	T1	14.611	0.162	2.907	0.414	0.012	1.513	0.275	67.106	5.939	0.748	.....	93.689
Manzanilla	MT2	T2	14.418	0.154	2.714	0.383	0.011	1.544	0.275	63.995	6.304	0.751	.....	90.547
Manzanilla	MT2	T2	14.103	0.154	2.741	0.392	0.011	1.547	0.271	64.472	6.368	0.753	.....	90.813
Manzanilla	MT2	T2	14.150	0.155	2.746	0.375	0.011	1.471	0.276	65.533	6.005	0.745	.....	91.467
Manzanilla	MT2	T2	14.179	0.158	2.843	0.407	0.012	1.514	0.272	65.221	6.146	0.767	.....	91.520
Manzanilla	MT3	T3	13.615	0.156	2.739	0.411	0.013	1.580	0.283	66.281	6.446	0.796	.....	92.320
Manzanilla	MT3	T3	13.419	0.154	2.723	0.404	0.012	1.575	0.279	66.487	6.481	0.796	.....	92.331
Manzanilla	MT3	T3	14.706	0.163	2.904	0.417	0.012	1.529	0.286	67.393	5.858	0.773	.....	94.042
Manzanilla	MT3	T3	14.689	0.162	2.915	0.427	0.012	1.516	0.276	67.780	5.874	0.775	.....	94.427
Hojiblanca	HT0	T0	13.872	0.136	2.062	0.325	0.012	0.895	0.243	66.845	5.399	1.028	.....	90.817
Hojiblanca	HT0	T0	14.110	0.160	2.270	0.362	0.013	0.907	0.268	66.834	5.275	1.018	.....	91.217
Hojiblanca	HT1	T1	13.460	0.154	2.199	0.358	0.013	0.967	0.270	67.453	5.542	1.058	.....	91.475
Hojiblanca	HT1	T1	13.417	0.154	2.200	0.357	0.013	0.962	0.269	67.296	5.525	1.055	.....	91.249
Hojiblanca	HT1	T1	12.923	0.159	2.231	0.353	0.013	0.984	0.277	68.253	5.526	1.007	.....	91.725
Hojiblanca	HT1	T1	13.301	0.162	2.278	0.358	0.013	0.989	0.279	67.616	5.539	1.017	.....	91.552
Hojiblanca	HT2	T2	13.266	0.155	2.228	0.359	0.013	0.999	0.270	66.439	5.855	1.030	.....	90.614
Hojiblanca	HT2	T2	13.892	0.155	2.164	0.345	0.013	0.987	0.268	65.632	5.739	1.012	.....	90.207
Hojiblanca	HT2	T2	13.319	0.159	2.263	0.365	0.014	0.998	0.268	66.735	5.824	1.054	.....	91.000
Hojiblanca	HT2	T2	12.653	0.157	2.257	0.365	0.013	1.007	0.276	67.226	5.922	1.070	.....	90.946
Hojiblanca	HT3	T3	13.108	0.157	2.266	0.370	0.013	1.033	0.270	68.029	6.228	1.101	.....	92.575
Hojiblanca	HT3	T3	13.148	0.157	2.271	0.368	0.013	1.037	0.278	68.104	6.232	1.102	.....	92.712
Hojiblanca	HT3	T3	13.326	0.159	2.270	0.370	0.013	1.025	0.279	67.913	6.085	1.085	.....	92.526
Hojiblanca	HT3	T3	13.293	0.159	2.268	0.370	0.014	1.019	0.276	67.643	6.058	1.081	.....	92.181

The central values for the different fatty acids were different. However, the differences were marked in only a few of them. For example, the average of C16:0 was 13.90 but the centered value (geometric mean) was 15.00; for C18:1c, the average was 66.74 whereas the centered value (geometric mean) was 72.61. Other acids gave closer values (C18:0 2.53 vs. 2.73; C16:1, 1.25 vs. 1.33) but only C21:0 led to the same result in both statistics (0.01), possibly because of its low proportion.

With respect to the variances of the log ratios (Table 2), the greatest values were observed for Ln (C21:0/C18:2n-6), although the ratio of C21:0 with respect to C18:3n-3, C18:1c, or C17:1 was also high. In relation to the ln of the ratios, it can be observed that the level of C16:0 is markedly higher that those of C17:0, C21:0, or C17:1. In fact, only C18:1c was in greater proportion than C16:0.

Table 2. Summary statistics for the classical statistics and cetered values when considered from a compositional data base.

	Variance ln(Xi/Xj)										
Xi\Xj	**C 16:0**	**C 17:0**	**C 18:0**	**C 20:0**	**C 21:0**	**C 16:1**	**C 17:1**	**C 18:1c**	**C 18:2n-6**	**C 18:3n-3**	**clr variances**
C 16:0		0.0005	0.0006	0.0017	0.0068	0.0017	0.0013	0.0008	0.0040	0.0027	0.0009
C 17:0	-4.5016		0.0001	0.0006	0.0052	0.0012	0.0007	0.0002	0.0036	0.0021	0.0003
C 18:0	-1.6183	2.8833		0.0007	0.0059	0.0017	0.0011	0.0003	0.0041	0.0026	0.0007
C 20:0	-3.567	0.9346	-1.9487		0.0039	0.0014	0.0012	0.0006	0.0041	0.0023	0.0006
C 21:0	-7.0854	-2.5838	-5.4671	-3.518		0.0051	0.0054	0.0047	0.0089	0.0067	0.0042
C 16:1	-2.2379	2.2637	-0.6196	1.3291	4.8475		0.0004	0.0009	0.0010	0.0007	0.0004
C 17:1	-3.9472	0.5543	-2.3289	-0.38	3.1382	-1.709		0.0005	0.0019	0.0009	0.0003
C 18:1c	1.5368	6.0384	3.1551	5.1038	8.6222	3.7747	5.484		0.0031	0.0018	0.0002
C 18:2n-6	-0.8462	3.6554	0.7721	2.7208	6.2392	1.3917	3.1011	-2.383		0.0010	0.0021
C 18:3n-3	-2.9298	1.5718	-1.3115	0.6372	4.1556	-0.692	1.0175	-4.467	-2.0836		0.0010
	mean ln(Xi/Xj)										0.0106 **Total Variance**

The tests performed in relation to the normality of the data (Table 3) showed that the determinations followed a normal distribution in all cases because the p values were always below 0.05.

Table 3. Results of the tests of normality applied to the values of the fatty acid composition data base.

	Anderson-Darling		**Cramer-von Mises**		**Watson**	
	A^2*	**p**	**W^2***	**p**	**U^2***	**p**
alr(C 16:0,C 18:3n-3)	2.2147	<0.01	0.3517	<0.01	0.3517	<0.01
alr(C 17:0,C 18:3n-3)	2.0185	<0.01	0.3373	<0.01	0.3373	<0.01
alr(C 18:0,C 18:3n-3)	3.3922	<0.01	0.5407	<0.01	0.5406	<0.01
alr(C 20:0,C 18:3n-3)	3.1168	<0.01	0.5033	<0.01	0.5032	<0.01
alr(C 21:0,C 18:3n-3)	0.9981	[0.025, 0.01]	0.1639	[0.025, 0.01]	0.1587	[0.025, 0.01]
alr(C 16:1,C 18:3n-3)	4.2648	<0.01	0.6627	<0.01	0.6627	<0.01
alr(C 17:1,C 18:3n-3)	2.7947	<0.01	0.4545	<0.01	0.4545	<0.01
alr(C 18:1c,C 18:3n-3)	2.5242	<0.01	0.4128	<0.01	0.4127	<0.01
alr(C 18:2n-6,C 18:3n-3)	3.1729	<0.01	0.5112	<0.01	0.5112	<0.01

The data were also subjected to exploratory data analysis by biplot (Figure 1 left). The highest variance with respect to their respective geometric means was observed for C16:1 and C18:3n-3. The link between these variables also provides information on the relative variability of ln(C16:1/C18:3n-3), which is the higest among all possible links. Furthermore, they are negatively correlated because the angle that they form is close to 180º. In the plot, the sample HT0 is clearly separate from the rest of the samples for the same cultivar. Then, this sample could be an aberrant point. Another interesting feature of the covariance biplot is that the variables are close to being collinear, which means that the composition plots along a compositional line (has one dimensional variability). In this case it could be principal component 1 (PC1). In the form biplot (which preserve the distances among samples) it is clearly observed (Figure 1right) that PC1 has distinguished between cultivars while CP2 represents possible effects of treatments. According to the distances within cultivars, it can be stated that the effects of treatments was limited and slighly higher in the case of Manzanilla. In this case, however, the distance of HT0 from the rest of Hojiblanca samples is lower, in agreement with the results of the tests of normality.

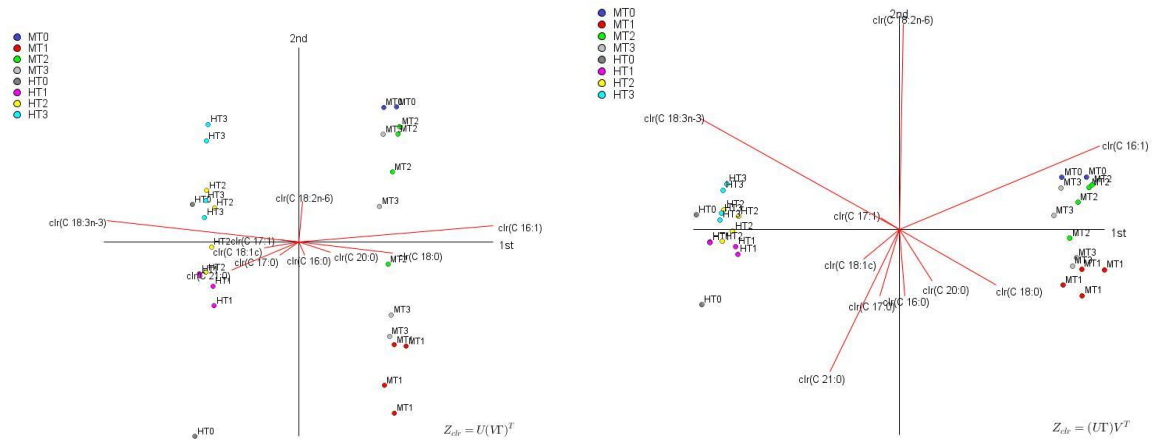


Figure1. Covariance biplot (left) and form biplot (right) of the compositional data of Manzanilla and Hojiblanca cultivars according to processing steps.

The linear relationship among the variables for the interaction treatments vs. cultivar was also studied by PCA in the simplex (Figure 2). As can be observed, the samples for each cultivar are closely grouped according to cultivars and separated between them but linked by CP1. The data were also subjected to an exploratory analysis in coordinates. Its associated orthonormal coordinates, being a vector of real variables, can be treated with the existing battery of conventional descriptive analysis.

In this case, there was no previous information to have in mind when applying the sequential binary partition. The first approach was to consider saturated acids vs. unsaturated ones. Then, the succesive partitions within each previous group were made according to molecular weight or saturation degree.

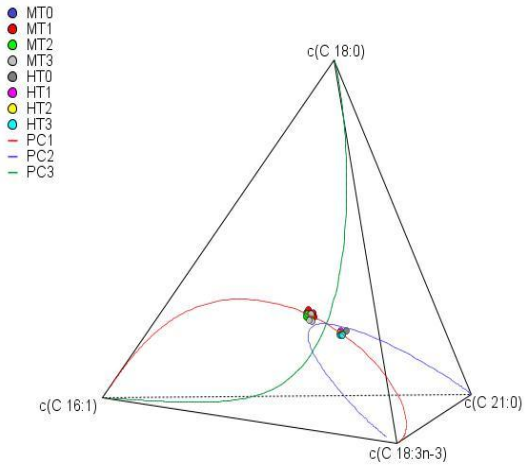


Figure 2. Ternary principal component analysis of data from the fatty acid composition. Samples correspond to the interaction treatments vs. cultivars.

The balance-dendograma shows (Figure 3) that, regardless of cultivar, the proportion of unsaturated acids is higher with respect to the unsaturated ones (the situation of the first balance is displaced to the left). The second balance shows that the proportion of lower molecuar weight fatty acids (C16:0, C17:0, and C16:0) is higher than those of higher number of carbons (C20:0 and C21:0). Considering the successive balances, one may deduce the relative abundance of the different groups of fatty acids and each of them with respect to the others. With respect to the variances, it is observed that, within cultivars, they are relatively low.

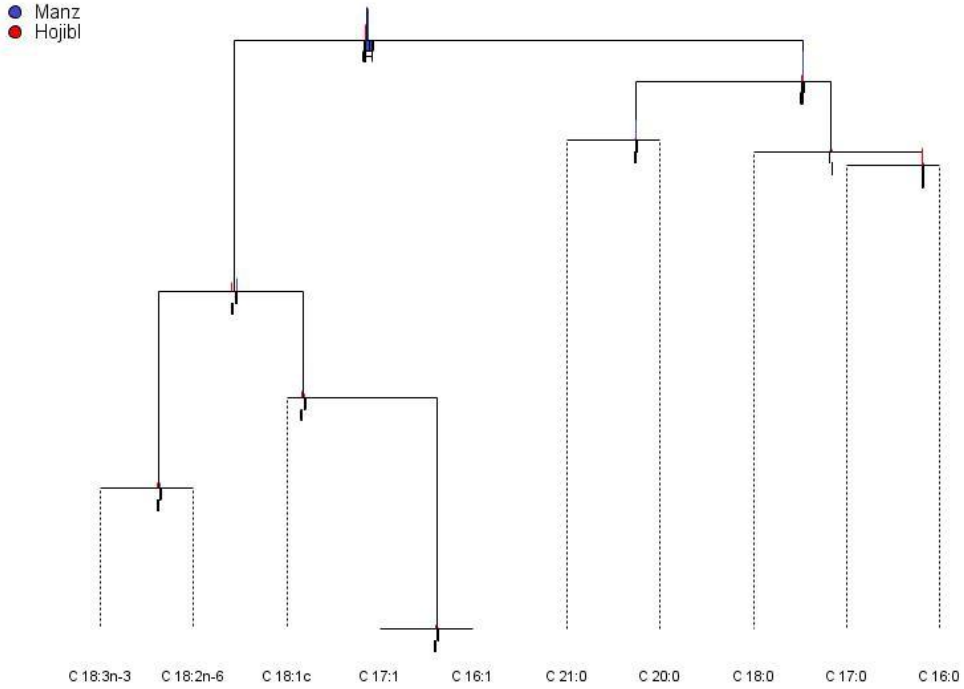


Figure 3. Balance dendogram of data from the fatty acid composition. Samples are grouped according to cultivars.

The balances were also used to study the effect of treatments on the fatty acid composition according to cultivars. A comparison between the results obtained applying the multivariate techniques directly to the raw data and to the ilr coordinates was also made. A first approach was the application of cluster analysis. The results are shown in Figure 4.

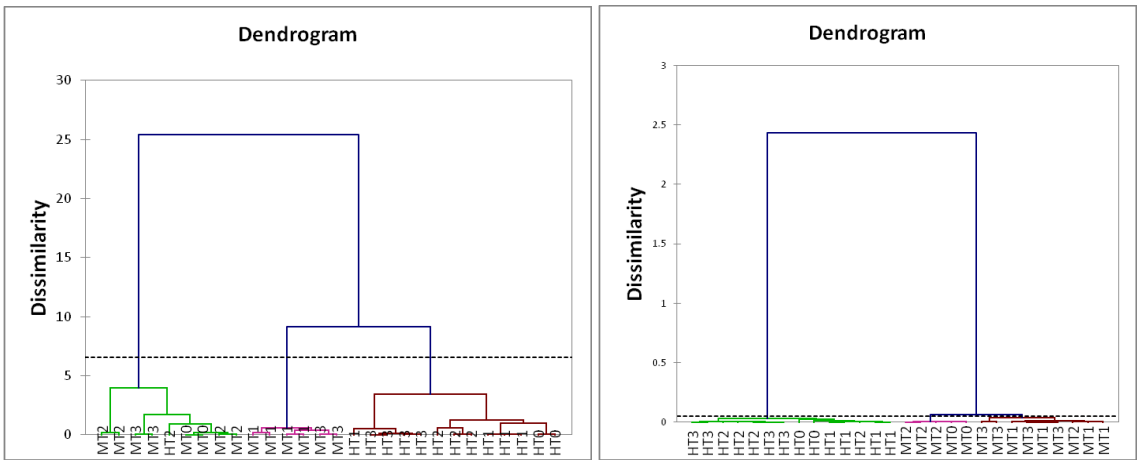


Figure 4. Results of the cluster analysis applied to the raw data (left) and ilr coordinates (right).

In the first case (raw data) there was a clear separation of three groups. Furthermore, dissimilarity observed within Manzanilla was markedly high and some of their treatments were closer to those of Hojiblanca. However, when the analysis was performed with ilr balances, great dissimilarity between cultivars (clear separation of cultivar) was observed without any wrong assignation of samples. In addition, dissimilarities within cultivar were also very limited. Hence, the application of clustering to ilr balances led to more realistic results. The application of robust discriminant analysis to data (Figure 5) also produced a clear segregation between cultivars.

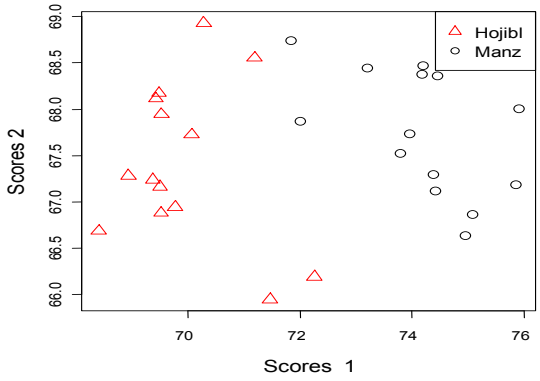


Figure 5. Fisher discriminate scores of two cultivars. Results from the DA, using robComposition software.

Data were also subjected to factor analysis, using a principal factor as extraction method. The Cronbach's alpha for raw data was 0.538 while in case of ilr coordinates the value was 0.827.

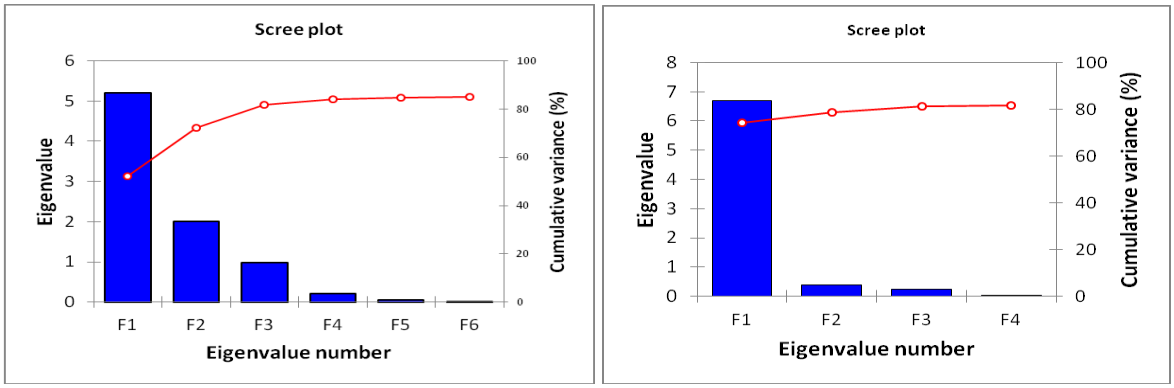


Figure 6. Scree plot for the succesive egenvalues

As deduced from the figures, in the case of raw data, the number of eigenvalues to be extracted (> 1) was three while in the second case, there was only one. According to these results, the variability of the raw data would be expressed by three hidden variables while in the second case only one is enough. The equations were:

Raw data

F1=-0.12C16:0+0.12C17:0+1.46C18:0-0.77C20:0+0.12C:21:0+1.61C16:1-0.12C17:1-0.09C18:1c-0.26C18:2n-6+1.14C18:3n-3

F1=0.33C16:0-0.41C17:0-0.35C18:0-0.49C20:0-0.04C:21:0-1.26C16:1-0.17C17:1+0.10C18:1c+0.51C18:2n-6-1.79C18:3n-3

F1=-0.09C16:0+1.06C17:0-4.36C18:0-0.23C20:0-0.47C:21:0+4.67C16:1-0.74C17:1+0.39C18:1c+0.65C18:2n-6+0.54C18:3n-3

-ilr coordinates:

F1=-0.278ilr1-0.009ilr2+1.544ilr3-0.453ilr4+0.373ilr5+1.266ilr6-1.441ilr7+2.957ilr8-0.248ilr9

This is clearly reflected in the projection of the variables (raw and ilr coordinates) onto the first two factors (Figure 7).

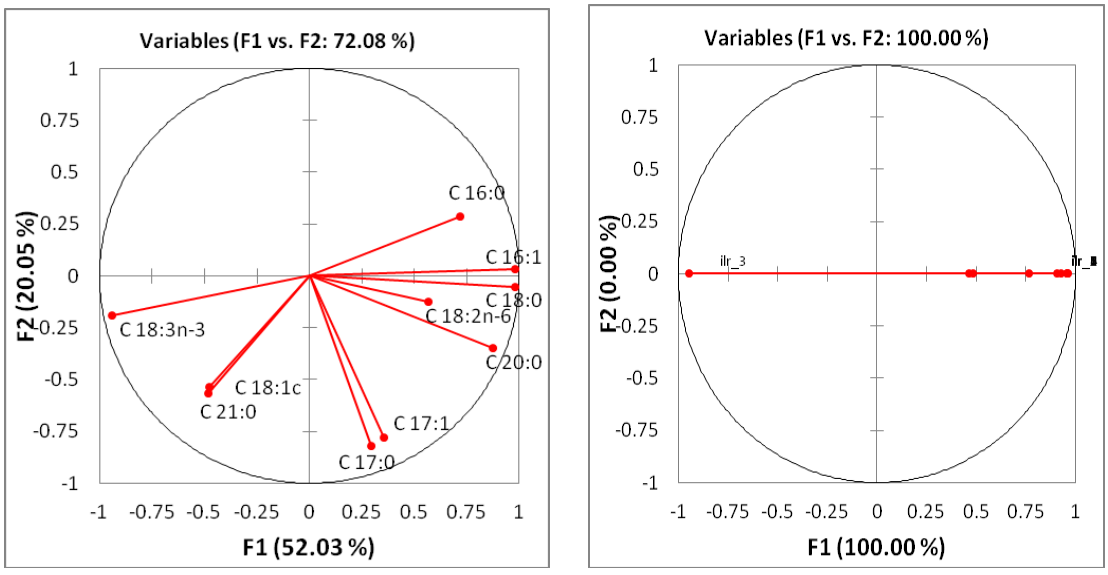


Figure 7. Projection of the original variables (raw data , on the left, and ilr coordinates, on the right) onto the first two factors.

The analysis made on the values considered as compositional data shows important differences with respect to that obtained when using the raw data. It is importat to emphasize that the linear relationship among the data was not detected when the conventional multivariate analysis was applied to the raw data.

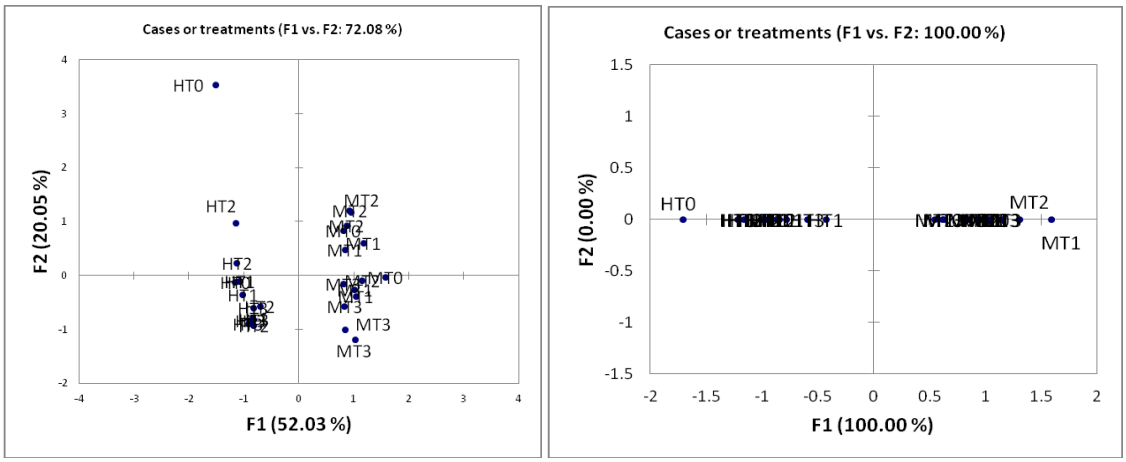


Figure 8. Pojection of cases (samples) on the plane of the first two factors.

In addition, the factor analysis led to a good segregation between cultivars (Figure 8), regardless of the type of data. This shows that in this aspect their respective results were comparable. Furthermore, the apparent aberrant sample (HT0) was detected in both cases and the distance of it with respect to the rest of samples from Hojibalnca was even higher using raw data. However, the images showed clear discrepancies. Particularly, the graphic obtained with the ilr coordinates reflects the linear relationship among the fatty acid compositions of the fat from both cultivars. Apparently, these fats differ in the relatioship among fatty acids while changes due to treatments do not alter them. Hence, results confims the original presumption of fat stability during the green Spanish-style table olive elaboration because this produces only limited modifications of the fatty acid composition.

The results also show that the application of standard multivariate analysis to raw data (fatty acid expressed as percentages) may lead to marked errors in certain circunstances. Furthermore, if one considers that all the legislation on olive oil classification is based on percentages, it is apparent that, at least some of the limits (percentages) should be expressed in other parameters (possible log ratio between individual fatty accids or groups of them). In summury, the results deduced from this work have shown that all the litterature on fatty acid composition of oils, and fats in general, should be subjected to a deeper revision in the future.

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.

Aparicio, R., Harwood, J. (2003). *Manual del Aceite de Oliva*. AVM Ediciones y Mundi Prensa. Madrid.

Comas-Cufí M, Thió-Henestrosa S. (2011) 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.

Commission Regulation (EEC) No 2568/91 of July 1991. On the characteristics of olive and olive pomace oils and on ther analystical methods. *Official Journal of the European Communities* L 248.

Fernández Díez, M.J., de Castro Ramos, R., Garrido Fernández, A., González Cancho, F., González Pellissó, F., Nosti Vega, M., Heredia Moreno, A., Mínguez Mosquera, M.I., Rejano Navarro, L., Durán Quintana, M.C., Sánchez Roldán, F., García García, P. and de Castro, A. (1985). "Biotecnología de la aceituna de mesa". CSIC, Madrid, Spain. CSIC, Ed. ISBN 84-00-06018-0.

Gambacorta, G., Storelli, M., Liuzzi, V., La Notte, E. (2002). Olive oil identity determined by a methodological and statistical procedure base don evaluationg the glyceridic fraction. *Ital. J. Food Sci*. 14, 59-64

Garrido-Fernández, A., Fernández-Díez, M.J., & Adams, R.M. (1997). *Table olives. Production and Processing*. London: Chapman & Halls.

IOOC (International Olive Oil Council) (2011). Table olive balances 2011. www.internationaloliveoil.org/ Last access April 2013.

Lee, D.S., Noh, B.S., Bae, S.Y., Kim, K. (1998). Characterization of fatty acids composition in vegetable oils by gas chromatography en chemometrics. *Analytica Chemica Acta* 358, 163-175.

López López, A., Rodríguez Gómez, F., Cortés Delgado, A., Montaño, A., Garrido Fernández, A. (2009). Influence of ripe table olive processing on oil characteristics and composition as determined by chemometrics. *J. Agric. Food Chem*. 57, 8973-8981.

López López, A., Rodríguez Gómez, F., Cortés Delgado, A., Montaño, A., Garrido Fernández, A. (2010). Effect of previous storage of ripe olives on the oil composition of fruits. *J.Am.Oil Chem. Soc*.87, 705-714.

López, A., Montaño, A., García, P., Garrido Fernández, A. (2006). Fatty acid profile of table olives and its multivariate characterization using unsupervised (PCA) and supervised (DA) chemometrics. *J. Agric. Food Chem*. 54, 6747-6753.

Martínez, J.M., Muñoz, E., Alba, J. and Lanzón, A. (1975). Informe sobre la utilización del analizador de rendimientos "Abencor". Grasas y Aceites 26, 379-385.

Pawlowsky-Glahn, V. and J. J. Egozcue (2011). Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment (SERRA) 15 (5), 384-398.

Templ, M., K. Hron, and 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. John Wiley & Sons, Chichester (UK), pp. 341-355.

Vlahov, G. (1996). The structure of triglycerides of monovarietal olive oils: a ^{13}C-NMR comparative study. Fett/Lipid 98, 203-205.

Compositional analysis of a mixture distribution with application to categorical modelling

MONIQUE GRAF[1] and DESISLAVA NEDYALKOVA[2]

[1]Institut de Statistique - Université de Neuchâtel, Switzerland
and Elpacos Statistics, Switzerland monique.p.n.graf@bluewin.ch
[2]Statistical Methods Unit - Federal Statistical Office - Switzerland

1 Introduction

In econometrics and other areas the necessity to account for lack of symmetry in the distribution of a quantity of interest is widely recognized. There are many different approaches to the problem. One way is to modify the usual normal or Student distribution by multiplying the density by a (normalized) cumulative distribution, giving rise to a skewed normal or skewed Student distribution, see Azzalini and Genton (2008). Another way is to model the tails of the distribution separately as e.g. Van Kerm (2007). A third approach that will be utilized here, is to fit a flexible enough parametric distribution that will account for lack of symmetry in the distribution. We concentrate on continuous distributions, but the method could be easily applied to discrete distributions as well.
Many probability distributions can be represented as compound distributions, see e.g. Johnson et al. (1995) and Kleiber and Kotz (2003). Consider some vector of parameters $\boldsymbol{\theta}$ as random with a probability density $h(\boldsymbol{\theta})$. Let $\mathbf{y}$ be the vector of observations. Specify a conditional distribution for the observations by the density $g(\mathbf{y}|\boldsymbol{\theta})$. The compound distribution is the marginal distribution of $\mathbf{y}$. We shall consider a positive range for the components of $\boldsymbol{\theta}$. Thus the marginal density of $\mathbf{y}$ is

$$f(\mathbf{y}) = \int_0^\infty g(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})\, d\boldsymbol{\theta}.$$

The densities f, g and h may depend on some other deterministic parameters.

The compound representation of f can be viewed as an infinite mixture of the densities $g(\mathbf{y}|\boldsymbol{\theta})$, the weighting scheme being given by the density $h(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}$. Mixture distributions are widely used in applications. Redner and Walker (1984) describe the maximum likelihood estimation through the EM algorithm of Dempster et al. (1977). For instance Reynolds and Templin (2004) consider mixed stock analysis in wildlife management and test for mixture homogeneity across samples. Starting from a baseline estimate from each population, the problem is to determine the proportions from each population in an observed mixture sample. Then the authors make a comparison of the mixture weights composition between two locations and tests of mixture equality by 3 different methods.
We take the problem the other way round: Having a priori populations determined by certain characteristics, we want to compare the distribution of a continuous measurement between these populations and the total population. Our approach can be summarized as follows: Suppose we have a distribution possessing the compounding property. This property can be interpreted as the consequence of a mixture of populations. The latent parameter $\boldsymbol{\theta}$ then gives a measure of discrepancy between populations. If we define a partition of the domain of definition of $\boldsymbol{\theta}$ into L parts $D_1, ..., D_L$, we can represent the above density f as a finite mixture of densities:

$$f(\mathbf{y}) = \sum_{\ell=1}^{L} \int_{D_\ell} g(\mathbf{y}|\boldsymbol{\theta})h(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \sum_{\ell=1}^{L} p_\ell\, f(\mathbf{y}|\boldsymbol{\theta} \in D_\ell). \tag{1}$$

The conditional densities $f_\ell(\mathbf{y}) = f(\mathbf{y}|\boldsymbol{\theta} \in D_\ell)$ will be called the *component densities* in the mixture representation (1) of $f(\mathbf{y})$. The weights $p_\ell = \Pr(\boldsymbol{\theta} \in D_\ell)$ of the component densities represent the probability that the underlying vector of parameters $\boldsymbol{\theta}$ belongs to domain D_ℓ, and together form a composition $\mathbf{p}$ (Aitchison, 2003). Of course the specification of the overall distribution f determines the probabilities p_ℓ. We first fit f to the overall population and choose a partition $D_1, ..., D_L$ of the

domain of variation of the latent variable $\boldsymbol{\theta}$. This gives rise to natural component densities f_ℓ. Thus for the overall population, we know the propabilities p_ℓ, but if we let the composition $\mathbf{p}$ vary freely, the representation (1) permits to embed the density f in a larger class of densities. The original distribution is thus used to define convenient component densities, then the mixture probabilities are modeled with auxiliary information that define subgroups. Our datasets stem from surveys with inequal probability sampling. Estimation proceeds with pseudo-maximum likelihood, that is the log-likelihood is weighted with the survey weights.

Our examples use the European Union Statistics on Income and Living Conditions (EU-SILC 2006) data and the Swiss SILC 2009. Graf et al. (2011) introduce the compound representation after fitting the income distribution with a GB2 distribution (McDonald, 1984). Keeping the component densities f_ℓ fixed, a linear model for predicting logratios of parts in $\mathbf{p}$ is introduced, thus modifying the original model. The necessary computations are available in in R-package 'GB2' Graf and Nedyalkova (2012) and the performance of the procedure is investigated in Graf and Nedyalkova (2011).
For each country, we fit model (1) in which $\mathbf{p}$ is predicted with household categories. Comparisons across countries are processed using robust principal component analysis for compositional data as implemented in Templ et al. (2011). The same procedure is applied to the Swiss SILC 2009 survey for which design-based variance computation is computable. Interesting balances (Egozcue and Pawlowsky-Glahn, 2005) are presented. Computations are processed using the open source program R (R Development Core Team, 2008).

The paper is organized as follows. In Section 2 we give a theoretical justification for the decomposition of a particular coumpound distribution, the generalized beta distribution of the second kind (GB2). Two different decompositions are presented: with respect to the right or the left tail of the distribution. In Section 3 we explain how the compounding property of the GB2 can be used in a survey context and in the context of small sub-populations. We define two models, with or without auxiliary information. The (pseudo)-loglikelihood, using the survey weights is derived and the method of estimation is presented. Analyses with the SILC data are presented in Section 4 and followed by a discussion in Section 5.

2 Decomposition of the GB2 distribution

2.1 The GB2 distribution

The GB2 distribution has been proposed by McDonald (1984) and has proven to give excellent fit to income distributions, see also McDonald and Butler (1987); McDonald and Xu (1995); McDonald and Ransom (2008); Jenkins (2008, 2009); Kleiber and Kotz (2003). Let us recall that the probability density of the GB2 with parameters a, b, p, q is given by:

$$f(x; a, b, p, q) = \frac{a}{b\, B(p, q)} \frac{(x/b)^{ap-1}}{(1 + (x/b)^a)^{p+q}} \tag{2}$$

with $a, b, p, q > 0$.
The GB2 parameters a, b, p, q need a large sample size (a few thousands) in order to be estimated with an acceptable precision. The GB2 model is thus hardly applicable to domains, even of moderate size. The maximum pseudo-loglikelihood used in Graf and Nedyalkova (2013) follows the lines of White (1982) and more specifically in a survey context those of Pfeffermann and Sverchkov (2003). The same method is applied by Biewen and Jenkins (2006). Aitchison (1975, 1990) gives methods to compare different distribution fits. An R package by Lumley (2010) permits to incorporate the survey design into the procedure. Results with the GB2 fit without mixture decomposition for EU-SILC data are presented in Graf and Nedyalkova (2013).
The compounding property of the GB2 distribution will allow us to exploit the model fitted at the overall level, also for small sub-populations. The idea behind is that the population consists of heterogeneous groups with respect to the scale of income and that this heterogeneity is well represented by the GB2. The aim is to set up a model that estimates the heterogeneity of subgroups and is consistent

with the overall fit. Once the distribution of incomes in the subgroup is determined, any subgroup characteristic (e.g. an indicator of poverty and social exclusion) can be computed.

The GB2 distribution can be expressed as an infinite mixture of distributions with varying scale parameters, that is as a compound distribution, (see Kleiber and Kotz, 2003, Table 6.1). Thus, as quoted by Kleiber and Kotz (2003), the GB2 distribution and its subfamilies can be given a theoretical justification as a representation of incomes arising from a heterogeneous population of income receivers. The compounding property will be used to derive a decomposition of the GB2 into a finite mixture of components.

Starting with a generalized gamma distribution $GG(a, \theta, p)$ with scale parameter θ, we obtain the GB2 distribution by assigning an inverse generalized gamma distribution $InvGG(a, b, q)$ to θ (see, e.g. Johnson et al., 1995).

The density $g(.; a, \theta, p)$ of $GG(a, \theta, p)$ is given by

$$g(x; a, \theta, p) = \frac{a}{\theta\,\Gamma(p)}(x/\theta)^{ap-1}\exp -(x/\theta)^a \tag{3}$$

and the density $h(.; a, b, q)$ of the distribution $InvGG(a, b, q)$ is

$$h(\theta; a, b, q) = \frac{a}{b\,\Gamma(q)}(\theta/b)^{-aq-1}\exp -(\theta/b)^{-a} \tag{4}$$

The GB2 density is obtained by integration over θ:

$$f(x; a, b, p, q) = \int_0^\infty h(\theta; a, b, q)\, g(x; a, \theta, p)\, d\theta \tag{5}$$

2.2 Decomposition with respect to the right or the left tail

Notice that the distribution of the random scale parameter θ does not depend on the shape parameter p governing the left tail. For this reason, we denote the decomposition in Equation (5) a "decomposition with respect to the right tail". We propose a similar "decomposition with respect to the left tail". It can be obtained using the following property of the GB2: Let $y = 1/x$ denote the inverse of the variable of interest x. Then y also follows a GB2 distribution and its density can be written as

$$f(y; a', b', p', q'),$$

where $a' = a$, $b' = b^{-1}$, $p' = q$ and $q' = p$ (see Kleiber and Kotz, 2003).

We have, using Equation (5):

$$f(y; a', b', p', q') = \int_0^\infty h(\theta; a', b', q')\, g(y; a', \theta, p')\, d\theta \tag{6}$$

By a change of variable, $(x = 1/y)$, in Equation (6), we obtain the left tail decomposition of the GB2 density in Equation (2):

$$f(x; a, b, p, q) = \int_0^\infty h(\theta; a, b^{-1}, p)\, (1/x^2) g(1/x; a, \theta, q)\, d\theta \tag{7}$$

When applied to an income variable, the decomposition with respect to the left tail emphasizes the variability of the poor and gives better results for the poverty indicators.

2.3 Right tail discretization

For simplicity, let us drop the explicit reference to the fixed parameters a, b, p, q in Equation (5). We propose to use the decomposition in the following way: Discretize the random scale parameter θ by partitioning its domain of integration into L intervals, with limits

$$\theta_0 = 0 < \theta_1 < \ldots < \theta_L = \infty.$$

Then write the GB2 density as a mixture:

$$\begin{aligned} f(x) &= \sum_{\ell=1}^{L} \int_{\theta_{L-\ell-1}}^{\theta_{L-\ell}} h(\theta) g(x,\theta)\, d\theta \\ &= \sum_{\ell=1}^{L} \left[\int_{\theta_{L-\ell-1}}^{\theta_{L-\ell}} h(\theta)\, d\theta \right] \frac{\int_{\theta_{L-\ell-1}}^{\theta_{L-\ell}} h(\theta) g(x,\theta)\, d\theta}{\int_{\theta_{L-\ell-1}}^{\theta_{L-\ell}} h(\theta)\, d\theta} = \sum_{\ell=1}^{L} p_{0\ell}\, f_\ell(x) \end{aligned} \tag{8}$$

The conditional density $f_\ell(x)$, given that the scale parameter is in $(\theta_{L-\ell-1}, \theta_{L-\ell})$, is defined by the fraction in Equation (8). The term in brackets is the probability $p_{0\ell}$, giving the weight of the density $f_\ell(x)$ in the mixture.

Evaluation of $f_\ell(x)$ and $p_{0\ell}$

With $u = (\theta/b)^{-a}$, the integration bounds are changed to

$$u_\ell = (\theta_{L-\ell}/b)^{-a}, \ell = 0, ..., L, \qquad (u_{\ell-1} < u_\ell).$$

Denoting by $P(\cdot, q)$ the cumulative distribution function of the standard gamma distribution with shape parameter q, we obtain

$$p_{0\ell} = P(u_\ell, q) - P(u_{\ell-1}, q) \tag{9}$$

In practice, the $p_{0\ell}$ are chosen and determine the u_ℓ.
Set $t = (x/b)^a + 1$. The density of the ℓ-th component in the mixture is given by:

$$f_\ell(x) = f(x) \frac{P(tu_\ell, p+q) - P(tu_{\ell-1}, p+q)}{P(u_\ell, q) - P(u_{\ell-1}, q)}, \tag{10}$$

where $f(x)$ is the GB2 density in Equation (2).

2.4 Left tail discretization

Knowing that the inverse of a GB2 random variable also follows a GB2 distribution, the principle is to apply the right tail discretization to the inverse y of the variable of interest and obtain a new decomposition in the original income scale by a change of variables $x = 1/y$.
For the inverse variable y with GB2 parameters $a' = a$, $b' = 1/b$, $p' = q$ and $q' = p$, we have: $u' = (\theta'/b')^{-a} = (\theta^{-1}/b^{-1})^{-a} = (\theta/b)^a$ and

$$u'_\ell = (\theta_\ell/b)^a, \ell = 0, ..., L, \qquad (u'_{\ell-1} < u'_\ell).$$

Knowing that $q' = p$, we see that u'_ℓ is determined by:

$$\tilde{p}_{0\ell} = P(u'_\ell, p) - P(u'_{\ell-1}, p).$$

With $t' = (y/b')^{a'} + 1 = (x/b)^{-a} + 1$, and changing to the variable $x = 1/y$, we obtain new component densities $\tilde{f}_\ell(x)$:

$$\tilde{f}_\ell(x) = f(x) \frac{P(t'u'_\ell, p+q) - P(t'u'_{\ell-1}, p+q)}{P(u'_\ell, p) - P(u'_{\ell-1}, p)} \tag{11}$$

Finally we have:

$$f(x) = \sum_{\ell=1}^{L} \tilde{p}_{0\ell}\, \tilde{f}_\ell(x) \tag{12}$$

Notice that in this representation, densities with more mass towards zero have a smaller index. Now, we can fit the compound GB2 distribution using this new decomposition of the GB2 density function. In the GB2 package, the factors of the GB2 density in Equations (10) and (11) are called "Gamma factors" and are obtainable through the function 'fg.cgb2'.

Figure 1 shows the right and left tail decomposition of the GB2 distribution for the equivalized income (Austrian EU-SILC sample 2006), with $p_\ell = \tilde{p}_\ell = 1/3$, $\ell = 1, 2, 3$. One sees clearly that the very poor are totally in f_1 for the left tail decomposition (bottom pane), but are scattered between all 3 components in the right tail decomposition (upper pane).

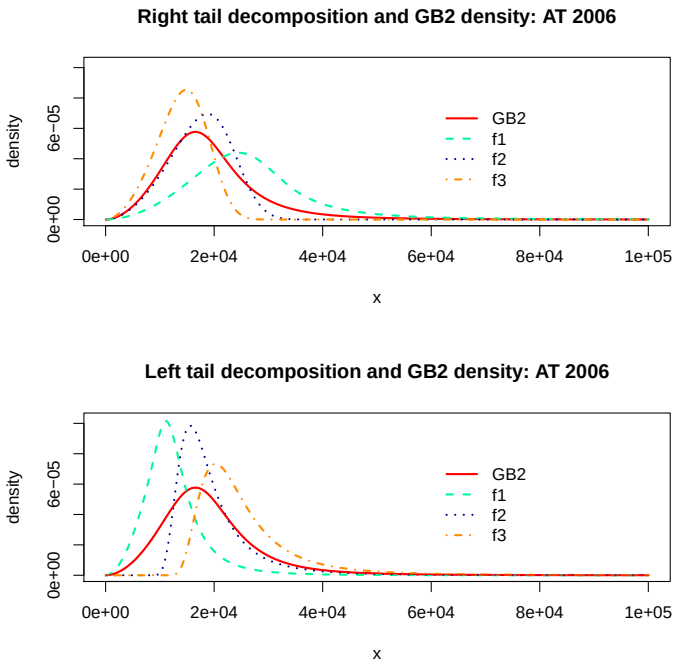


Figure 1: Right and left tail decomposition and the parent GB2 density and $p_{\ell 0} = 1/3, \ell = 1, 2, 3$.

3 Models based on the decomposition

In our SILC examples, the GB2 parameters a, b, p, q are determined at the overall country level. See Graf and Nedyalkova (2013) for the estimation method, and Graf and Nedyalkova (2011) for the simulation results. Now, given a partition into L intervals for the scale parameter θ of incomes, we can define a new model for a subpopulation based on a mixture of the densities $f_\ell(.)$, given in Equation (8), or $\tilde{f}_\ell(.)$, given in Equation (12). In this model, the component densities f_ℓ of the mixture are fixed and the probabilities p_ℓ are re-fitted at the subpopulation level.

The initial GB2 fit $p_{0\ell}$ of p_ℓ, given by the term in brackets in Equation (8) or (12) will serve as starting values $p_\ell^{(0)}$. The estimation method is by maximum pseudo-likelihood (PML), as used for the GB2 fit (and is implemented in the R-package 'GB2', Graf and Nedyalkova, 2012). We can use the procedure in two ways:

1. Fit the p_ℓ on a subpopulation.
   It is assumed that we need a much smaller sample size for a good estimate of the probabilities p_ℓ than it was necessary for the estimation of the GB2 parameters.

2. Model the p_ℓ with auxiliary information.
   Auxiliary variables can be used to model the probabilities p_ℓ, without reference to the density $h(\cdot)$. In this way, heterogeneous population structures can be accounted for.

In both cases, an iterative algorithm is constructed.

3.1 Maximum pseudo-likelihood estimation

Let us write for simplicity the component densities as f_ℓ. The estimation method is similar for $\tilde{f}_\ell$. Let n be the sample size. The pseudo-loglikelihood is written as

$$\log L(p_1, ..., p_L) = \sum_{k=1}^{n} w_k \log \left(\sum_{\ell=1}^{L} p_\ell f_\ell(x_k) \right) \tag{13}$$

This formula can be written as:

$$\log L(p_1,...,p_L) = \sum_{k=1}^{n} w_k \log\left(f(x_k)\right) + \sum_{k=1}^{n} w_k \log\left(\sum_{\ell=1}^{L} p_\ell \frac{P(t_k u_\ell, p+q) - P(t_k u_{\ell-1}, p+q)}{p_{0\ell}}\right) \quad (14)$$

There are only $L-1$ parameters to estimate, because the probabilities p_ℓ sum to 1. Moreover, the p_ℓ must be positive. With these constraints in mind, we change the parameters p_ℓ, $\ell = 1,...,L$, to

$$v_\ell = \log(p_\ell/p_L), \qquad \ell = 1,...,L-1,$$

and maximize the likelihood expressed as a function of the v_ℓ. Variance estimation is done via the sandwich variance estimator. More details will be provided in the full paper.

3.2 Introduction of auxiliary variables

Let us model the probabilities p_ℓ with auxiliary variables. Let $\mathbf{z}_k$ be the vector of auxiliary information for unit k. This auxiliary information modifies the probabilities p_ℓ at the unit level. Let us denote by $p_{k,\ell}$ the weight of the density f_ℓ for unit k. For $\ell = 1,...,L-1$, we pose a linear model for $v_{k,\ell}$:

$$\log(p_{k,\ell}/p_{k,L}) = v_{k,\ell} = \sum_{i=1}^{I} \lambda_{\ell i} z_{ki} = \mathbf{z}_k^T \boldsymbol{\lambda}_{\boldsymbol{\ell}} \quad (15)$$

The likelihood is then maximized with respect of the $\boldsymbol{\lambda}_{\boldsymbol{\ell}}$.

3.3 Choice of partition

The number L of components f_ℓ can be chosen arbitrarily, but it may be reasonable to keep L small. If we choose $L = 3$ and $p_\ell^{(0)} = 1/3$, the components f_1, f_2, f_3 represent respectively the income distributions with small, medium and high scale parameters, that is with more mass to the left for f_1, more mass to the center for f_2 and more mass to the right for f_3, each having the same weight in the overall GB2 fit. For the data in the example , we have found that $L = 5$ components give better results. A systematic way to choose the partition has still to be developed.

A thorough presentation of income distributions together with inequality measurement can be found in Atkinson and Bourguignon (2000) and in Chotikapanich (2008). The EU-SILC survey and its main variable, the equivalized disposable income, is presented in Clémenceau and Museux (2007). The weighting scheme is found in Osier et al. (2006). Income comparisons on the basis of the EU-SILC survey have been proposed by Jäntti (2007). The novelty here is the use of the decomposition of the GB2 and its application to categorical modelling.

4 Income distribution per household composition category

In the present example, our aim is to evaluate the predictive power on the equivalized income distribution of three categories of households, those without children (no.child), with a single adult with children (sa.ch) and with two or more adults with children (ma.ch). A child is defined as being aged 14 years old or less. The auxiliary variables are the indicator variables of the 3 household categories. The computations are made at the person's level, and not the household which seems to be more sensible in this context.

We work with a left tail decomposition into 5 components and $p_{0\ell} = \frac{1}{5}, \ell = 1,...,5$, the performance in reproducing the empirical inequality and poverty indicators being better than with 3 components. Computations are made with the R-package 'GB2' (Graf and Nedyalkova, 2012) (except for the variance computation). Consider first the Swiss SILC survey 2009. Table 1 shows the estimated parameters $\hat{\lambda}_{i\ell}$, by group i and component $\ell = 1,...,4$ and the corresponding standard errors $s_{i\ell}$. Let us recall that the parameters are the logratios of the mixture probabilities of the first $L-1$ components to the probability of the last. We see that the household categories are significantly different on at least one component. The large variability ($s_{34} = 8.48$) of the share of components 4 and 5 for "nochild" is to

Table 1: Parameter estimates and corresponding standard errors, Swiss SILC survey 2009

group i	$\hat{\lambda}_{i1}$	$\hat{\lambda}_{i2}$	$\hat{\lambda}_{i3}$	$\hat{\lambda}_{i4}$	s_{i1}	s_{i2}	s_{i3}	s_{i4}
total	-0.46	-0.16	-0.02	-2.62	0.10	0.18	0.22	2.96
ma.ch	0.58	1.55	1.18	-2.11	0.34	0.28	0.52	2.95
nochild	-0.85	-1.06	-0.47	-3.50	0.11	0.34	0.30	8.48
sa.ch	4.45	3.47	3.94	-14.48	0.34	1.14	0.64	0.21

be noted. The corresponding mixture probabilities are given in Table 2. Top row is the result of a fitting a compound GB2 to the total sample. The main discrepancy with the GB2 is at component 4, which empties to the profit of the two adjacent components. The bottom three rows in Table 2 were obtained by fitting the compound distribution with the group indicators as auxiliary variables. Households with more than one adult and children are over represented on the components 2 and 3 and under represented on the "richer" components 4 and 5. The category "nochild" has a smaller share of component 2 but a much larger one on component 5 than "total". For the category "sa.ch", half of the probability mass is attributed to the "poorer" component 1. In Figure 2 (top), we see that the compound GB2 mode is in better agreement than the GB2 mode with the kernel estimated modal value. The kernel estimated density near zero is not compatible with the truncation of incomes at zero, so we do not take too much attention to the discrepancy with the GB2 model. The compound densities per household category are represented at Figure 2 (middle) They are in good agreement with the kernel density estimates (figure 2, bottom).

Table 2: Compound distribution fitted mixing probabilities parameters, total population and by household categories

	p_1	p_2	p_3	p_4	p_5
total	0.179	0.240	0.278	0.021	0.282
ma.ch	0.165	0.433	0.299	0.011	0.092
nochild	0.177	0.142	0.257	0.012	0.412
sa.ch	0.504	0.188	0.302	0.000	0.006

Now we consider the 26 countries participating to the EU-SILC survey in 2006. We start with the GB2 fit separately by country (Graf and Nedyalkova, 2013) and ignoring the auxiliary information on the household categories. We apply to each country the same method as for the Swiss data above. We first fit a GB2 distribution, considering all persons with a positive equivalized income. Then we fit a compound GB2 with 5 components and equal mixture probabilities under the GB2, considering again all persons. Thirdly, we fit a compound distribution to each household category. In the third case, it is possible to fit either a model with the household categories indicators as auxiliary variables, or to fit a model to each category separately. The second option has been taken.

Principal component analysis (PCA) is a convenient tool to illustrate multivariate data. Aitchison and Greenacre (2002) have adapted the PCA to compositional data. Essentially, it is a PCA on the centred logratios of parts: the logarithm of each part divided by the geometric mean of parts (Aitchison, 2003). Filzmoser et al. (2009); Filzmoser and Hron (2011) have developed a robust version of the method. The following analyses are processed with the R package robCompositions (Templ et al., 2011).

In order not to blur the comparison of the household categories with the added flexibility due to the five components mixture, in the PCA analysis, we focus on the discrepancy between the mixture probabilities of the household categories and the mixture fit without auxiliary variables ('total'), in other words: on the perturbation induced by the auxiliary information. We show only the robust compositional PCA (Figure 3). The biplot has been represented in three panes, separately per household category, in order to better visualize the points.

Plane (1,2) explains 84% of the variability. Those countries near the origin in the three panes do

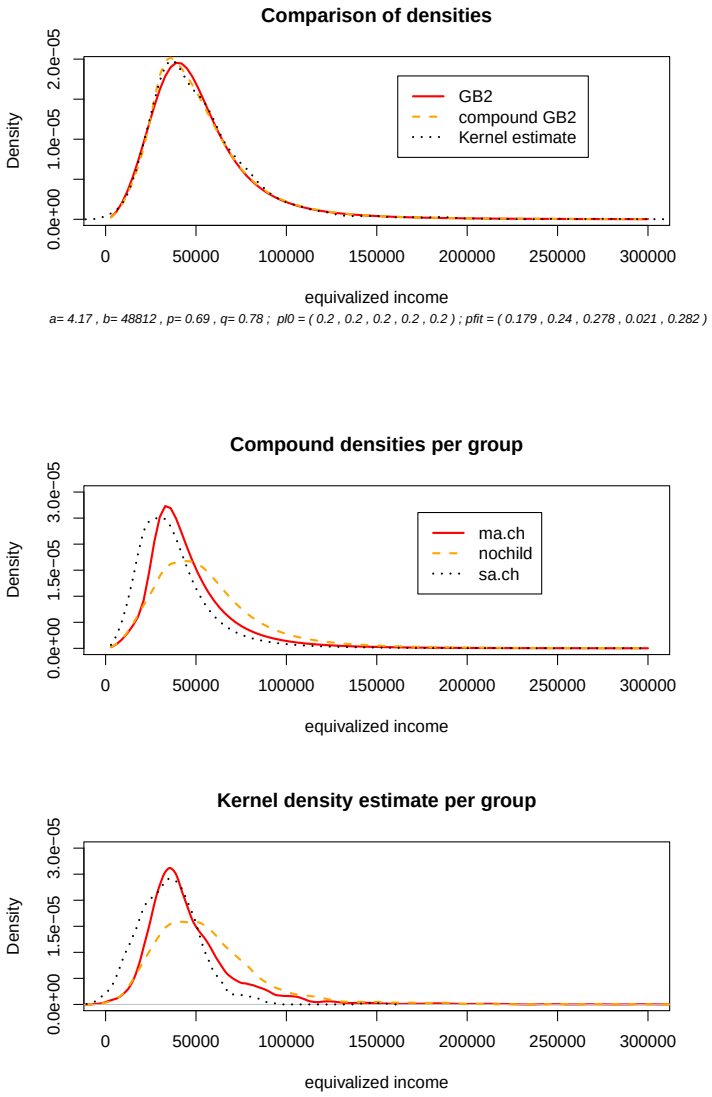


Figure 2: Swiss Silc data: all sample (top), compound fit with auxiliary variables (middle), Kernel estinate per household category (bottom). Equivalized income is in CHF.

not show a big difference between household categories. There is a tendency for 'ma.ch' to aggregate towards the middle component c3, and of 'no.child' to give more weight to c2 and c5. The category 'sa.ch' contribute manly to the component c2. Persons in households without children (nochild, bottom left) are clearly similar to the overall population, with a slight shift towards the 'richest' components c4 and c5. A slight shift towards components 3 and 4 can be observed for 'ma.ch'. Notice that the extreme point (IS on pane 'nochild' and DK and LV on pane 'sa.ch') have ill conditioned Hessians. It is worth noting that the third principal component has 9% explanatory power and is essentially related to c1 and the group 'sa.ch'. Thus, the groups clearly correspond to different mixtures.

5 Discussion

The advantage of GB2 modelling is that we often have a good first approximation of the distribution of interest. The decomposition presented in Section 2 provides a natural way of refining the fit. If for instance, we choose a left tail decomposition, then the asymmetry to the right of the distribution is essentially fixed by the components, but the mode and the asymmetry to the left are re-estimated. Further, the adjustment of the compound GB2 without using auxiliary information provides a benchmark

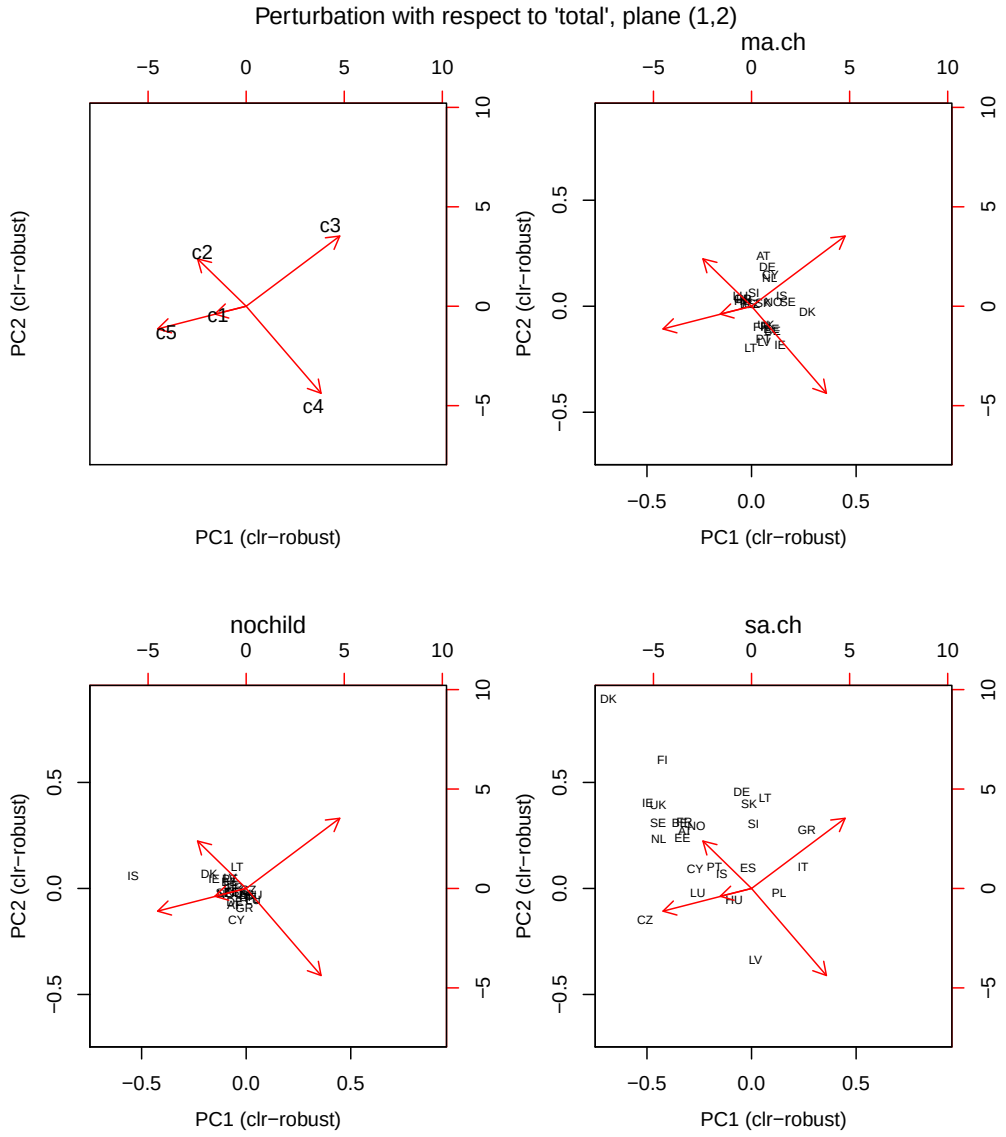


Figure 3: Compositional biplots showing the perturbation in the mixture with five components per household category

to the usefulness of incorporating this information into the model. It is not meaningful to substract probabilities. Instead, Aitchison (2003) coined the term 'perturbation' for the conditional probability of $\mathbf{p}_1$ given $\mathbf{p}_2$. When comparing different populations, each giving rise to a different GB2 fit, a useful tool is to consider the perturbation of the mixture probabilities with auxiliary information, with respect to the probabilities computed overall, i.e. with a constant model. This amounts to consider the GB2 adjustment as an ancillary information, the interest being in the change in distribution. In this way, despite the fact that the GB2 distribution are country-dependent, the results will be comparable across countries.

It could be possible to use Kullback-Leibler information along the lines of Aitchison (1990) to discriminate among models, but this has not been implemented for the while. Provided examples using categorical modelling gave interpretable results. The method can be applied to continuous auxiliary variables as well. In this case, the mixture probabilities $p_{k\ell}$ depend on unit k. The interpretation can be provided by compositional tools like compositional PCA.

Acknowledgements This research was partly supported by the FP7-SSH-2007-217322 AMELI Research Project.

References

Aitchison, J. (1975). Goodness of Prediction Fit. *Biometrika*, 62, 3:547–554. http://biomet.oxfordjournals.org/cgi/content/abstract/62/3/547.

Aitchison, J. (1990). On Coherence in Parametric Density Estimation. *Biometrika*, 77, 4:905–908. http://biomet.oxfordjournals.org/cgi/content/abstract/77/4/905.

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

Aitchison, J. and Greenacre, M. (2002). Biplots for compositional data. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 51(4):375–392.

Atkinson, A. B. and Bourguignon, F., editors (2000). *Handbook of Income Distribution*. Elsevier.

Azzalini, A. and Genton, M. (2008). Robust likelihood methods based on the skew-t and related distributions. *International Statistical Review*, 76(1):106–129.

Biewen, M. and Jenkins, S. P. (2006). Variance Estimation for Generalized Entropy and Atkinson Inequality Indices: the Complex Survey Data Case. *Oxford Bulletin of Economics and Statistics*, 68(3):371–383.

Chotikapanich, D., editor (2008). *Modeling Income Distributions and Lorenz Curves*. Springer: Economic Studies in Equality, Social Exclusion and Well-Being, Vol. 5.

Clémenceau, A. and Museux, J.-M. (2007). EU-SILC: an EU statistical instrument collecting cross national comparable data on income and living conditions and the measure of well being. In *Perspectives of improving economic welfare measurement in a changing Europe, 34th CEIES Seminar, Helsinki*.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*, 39, 1:1–38. http://www.jstor.org/pss/2984875.

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

Filzmoser, P. and Hron, K. (2011). Robust statistical analysis. In Pawlowski-Glahn, V. and Buccianti, A., editors, *Compositional Data Analysis*, pages 59–72. John Wiley & Sons, Ltd. http://dx.doi.org/10.1002/9781119976462.ch5.

Filzmoser, P., Hron, K., and Reimann, C. (2009). Principal component analysis for compositional data with outliers. *Environmetrics*, 20:621–632.

Graf, M. and Nedyalkova, D. (2011). Parametric estimation of income distributions and derived indicators using the GB2 distribution. In Hulliger, B., editor, *Report on the Simulation Results*, chapter 7.1. Deliverable 7.1 of the AMELI project.

Graf, M. and Nedyalkova, D. (2012). *GB2: Generalized Beta Distribution of the Second Kind: properties, likelihood, estimation.* R package version 1.1.

Graf, M. and Nedyalkova, D. (2013). Modeling of income and indicators of poverty and social exclusion using the Generalized Beta Distribution of the Second Kind. *Review of Income and Wealth*. to appear.

Graf, M., Nedyalkova, D., Münnich, R., Seger, J., and Zins, S. (2011). Parametric Estimation of Income Distributions and Indicators of Poverty and Social Exclusion. Technical report, WP2 - D2.1, FP7-SSH-2007-217322 AMELI Research Project. http://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/Ameli_Delivrables/AMELI-WP2-D2.1-20110409.pdf.

Jäntti, M. (2007). The EU-SILC in Comparative Income Distribution Research: Design and Definitions in International Perspective. In *Comparative EU Statistics on Income and Living Conditions: Issues and Challenges. Proceedings of the EU-Silc Conference (Helsinki, 6-8-November 2006)*. http://tilastokeskus.fi/eusilc/jantti.pdf.

Jenkins, S. P. (2008). Inequality and the GB2 Income Distribution. In *ISER Working Paper 2007-(Revised May 2008)*. 12. Colchester: University of Essex. http://www.iser.essex.ac.uk/pubs/workpaps/pdf/2007-12.pdf.

Jenkins, S. P. (2009). Distributionally-sensitive Inequality Indices and the GB2 Income Distribution. *Review of Income and Wealth*, 55(2):392–398.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions*, volume 2. New York: John Wiley, 2nd ed. edition.

Kleiber, C. and Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. John Wiley & Sons, Hoboken, NJ.

Lumley, T. (2010). survey: analysis of complex survey samples. R package version 3.23-3.

McDonald, J. (1984). Some Generalized Functions for the Size Distribution of Income. *Econometrica*, 52 (3):647–663.

McDonald, J. B. and Butler, R. J. (1987). Some Generalized Mixture Distributions with an Application to Unemployment Duration. *The Review of Economics and Statistics*, 69:232–240. http://www.jstor.org/pss/1927230.

McDonald, J. B. and Ransom, M. (2008). The generalized beta distribution as a model for the distribution of income: Estimation of related measures of inequality. In Chotikapanich, D., editor, *Modeling Income Distributions and Lorenz Curves*, volume 5 of *Economic Studies in Inequality, Social Exclusion and Well-Being*, pages 147–166. Springer New York. 10.1007/978-0-387-72796-7_8.

McDonald, J. B. and Xu, Y. J. (1995). A generalization of the beta distribution with applications. *Journal of Econometrics*, 66(1-2):133–152. Erratum: *Journal of Econometrics*,69, 427-428.

Osier, G., Museux, J.-M., Seoane, P., and Verma, V. (2006). Cross-sectional and longitudinal weighting for the EU-SILC rotational design. In *Methodology of Longitudinal Surveys, Essex, UK*.

Pfeffermann, D. and Sverchkov, M. Y. (2003). Fitting generalized linear model under informative sampling. In Skinner, C. and Chambers, R., editors, *Analysis of Survey Data*, pages 175–195. Wiley, New York, USA.

R Development Core Team (2008). R: A language and environment for statistical computing. ISBN 3-900051-07-0.

Redner, R. A. and Walker, F. W. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239.

Reynolds, J. H. and Templin, W. D. (2004). Comparing mixture estimates by parametric bootstrapping likelihood ratios. *Journal of Agricultural, Biological, and Environmental Statistics*, 9(1):57–74. DOI: 10.1198/1085711043145.

Templ, M., Hron, K., and Filzmoser, P. (2011). *robCompositions: an R-package for robust statistical analysis of compositional data*. John Wiley and Sons.

Van Kerm, P. (2007). Extreme Incomes and the Estimation of Poverty and Inequality Indicators from EU-SILC. *IRISS-C/I Working Paper*. http://iriss.ceps.lu/documents/irisswp69.pdf.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica*, 50:1–25.

Compositional data analysis of German national accounts

K. HRŮZOVÁ[12] and K. HRON[12]

[1]Department of Mathematical Analysis and Applications of Mathematics - Faculty of Science, Palacký University, Czech Republic, klara.hruzova@gmail.com
[2]Department of Geoinformatics - Faculty of Science, Palacký University, Czech Republic

1 Compositional data in economics

Compositional data frequently occur in biology, chemistry or geology but they also result from measurements in macro- and microeconomics. The aim of this paper is to provide an overview of possible presence of compositions in economics as well as to perform a detailed analysis for a concrete real-world problem - relative structure of gross value added in regions of Germany.

The typical example of economic compositional data are household expenditures. In this case an overall family budget is partitioned into particular outcomes (e.g. housing, food, education, culture, savings, etc.). Further examples, we can meet in practice, follow.

An investment portfolio can be regarded as a linear combination of assets. Composition here are weights of every asset involving in the portfolio (Härdle, Simar, 2012). We are interested in revenues of each asset and according to this information we want to optimize the portfolio in the sense of either the biggest profit or the lowest risk. It is clear if we want to gain utmost we have to accept more risky investments.

In banking there could be found an interesting problem - approving loans, credit scoring. There are many criterions, each with different scale and different values, that should satisfy each applicant. The question is how we could optimize this set of criterions and its values to minimize the relative amount of loans in delay or in other words non-paying loans. Consequently, classification issues are of particular interest here as well.

Many other problems can be found in stock markets. For example, we can consider percentage changes of the rate of assets as functional data (density functions) and use the correponding statistical tools for their further analysis. We can also regard the basis of the stock index, like the PX index in case of the Czech exchange market (with daily changes), as a composition consisting of fourteen the most tradable shares and analyse it using time series methodology.

Many other datasets appropriate for compositional data analysis can be found in regional economic statistics, national economic characteristics or in economic statistics of organisations and companies. It just depends which economic problems are of a particular interest.

2 National accounts

We will illustrate mentioned considerations with concrete real-world problem - we proceed to stuctural analysis of gross value added in German regions (recent data are available from the year 2009), see (Landesbetrieb für Statistik, 2012) for details. In order to undestand the underlying economic origin of the compositional data, we start with a definition of national accounts and gross value added in general and then we move to our concrete analysis.

National accounts are a measure of national economic activity (Lequiller, Blades, 2006). They present outputs, expenditure and income activities of the economic actors - households, corporations, governments - in economy.

Although there may be a variety of national accounts´ presentations that differ country by country. There are few main national accounts, namely:

- Current accounts
  - production accounts (describing the value of domestic output, balancing item is value added)

  - income accounts (showing primary and secondary income flows, balancing item is disposable income)
  - expenditure accounts (consumption or savings of disposable income, balancing item are savings)

- Capital accounts (recording net accumulation of nonfinancial assets, financing of the accumulation, balancing item is net lending/borrowing)
- Financial accounts (showing net acquisition of financial assets and the net incurrence of liabilities, balancing item is net change in financial position)
- Balance sheets (recording the stock of assets and liabilities at a particular point of time, balancing item is net worth)

2.1 Gross value added

The point of main interest in this paper is the gross value added (GVA) which is the balancing element of production accounts (Samuelson, Nordhaus, 1989). It is a measure of the value of goods and services produced in an area, industry or sector of an economy.

Gross domestic product represents three items: consumer spending, business investment and government spending. Consumer spending represents all expenditures in a nation by individual consumers. Business investments are all large purchases of equipment and facilities for production. Government spending is comprised of the expenditures on finished goods and services produced by the private sector.

Gross value added represents the output minus intermediate consumption. It is linked to gross domestic product (GDP) in the following sense:

$$\text{GVA} + \text{taxes on products} - \text{subsidies on products} = \text{GDP}.$$

GVA is used mainly for measuring gross regional domestic product and other measures of the output of entities smaller than the whole economy.

3 German economy

In order to proceed further to the real-world data with relative structure of GVA, we introduce Germany as one of the most highly developed and efficient industrial countries. With a population of 82 million inhabitants Germany is furthermore the largest and most important market in the European Union (EU). German GDP for the year 2012 is 2,643,900 million EUR and GVA is 2,364,510 million EUR. The proportion of service sector on GDP is around 71.1 %, further 28.1 % of industry and 0.8 % of agriculture. The price-adjusted GDP increased by 0.7 % compared with the previous year.

After the World War II Germany was divided into Federal Republic of Germany (west part controlled by France, the United Kingdom and the United States of America) and German Democratic Republic (part of Soviet Zone). This was the reason for different development of economies which tends to have an influence till today. With unification on 3rd October 1990 Germany's major effort is devoted to reconciling the economic systems of the two former countries.

The southern states, especially Bayern, Baden-Württemberg and Hessen, are economically stronger than the northern states. One of Germany's traditionally strongest (and at the same time oldest) economic regions is the Ruhr area in west, between Bonn and Dortmund. Even after the German reunification in 1990 the standard of living and annual income remains significantly higher in the former west German states. The modernisation and integration of the eastern German economy continues to be a long-term process scheduled to last until the year 2019.

4 German gross value added

The analysed dataset consists of the gross value added structure (2009) (Landesbetrieb für Statistik, 2012) in 16 regions, that is expressed in percentages of agriculture, production and services. Thus we have three-part composition, represented with a constant sum constraint 100%.

All the figures were plotted in statistical software R (available at www.r-project.org) and its package robCompositions (Templ et al., 2011).

At first we just plot the data matrix to see the relationship between parts of composition. In Figure 1 we can see an influence of spurious correlation in scatterplot of production vs. services. Nevertheless, from Figure 1 we can still conclude that the highest percentages of agriculture are in eastern Germany and they approach zero values in western Germany.
It is neccesary to note here that different colours distinguish the federal states of Germany in this sense (that correpond to natural geographical regions):

- yellow for Mecklenburg-Vorpommern, Sachsen and Thüringen,
- orange for Sachsen-Anhalt and Brandenburg,
- violet for Berlin, Schleswig-Holstein, Hamburg and Bremen,
- blue for Niedersachsen, Nordrhein-Westfalen, Hessen and Bayern,
- green for Rheinland-Pfalz, Baden-Württemberg and Saarland.

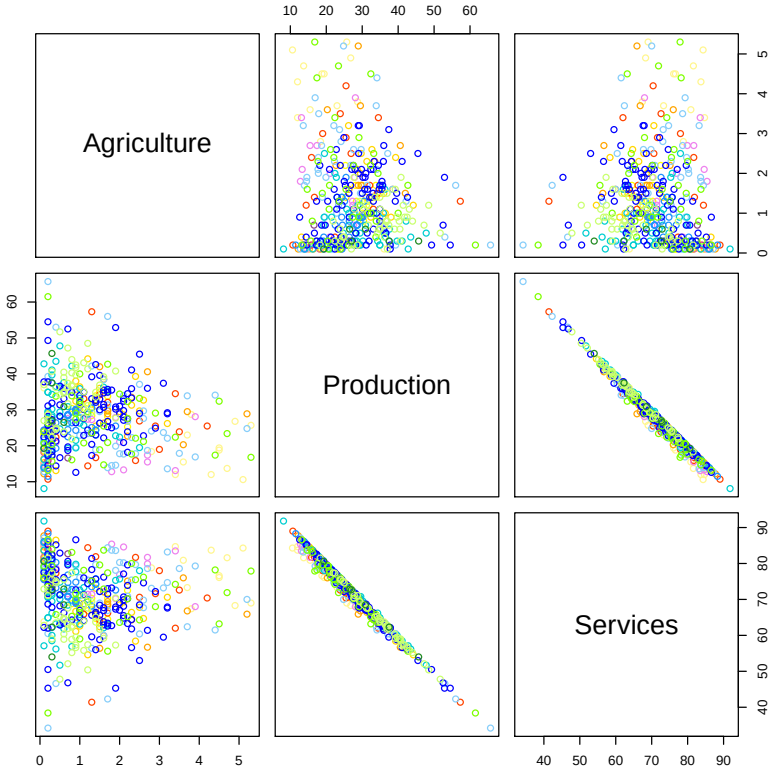


Figure 1: Data matrix with the original GVA structure data.

The three-part compositions can be plotted in ternary diagram. As we can see in Figure 2 (left) the data are clustered on the side between Production and Services; this means the Agriculture part

contains mostly small positive values. For better visualization we thus centered the compositions (in the Aitchison sense) and result is plotted in Figure 2 (right).

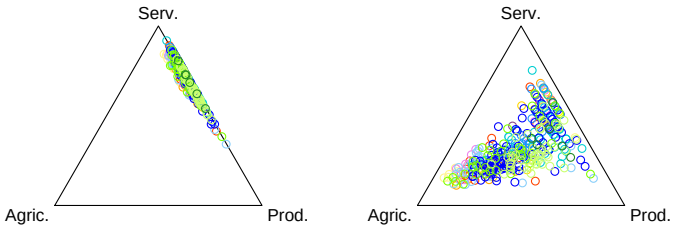


Figure 2: Ternary diagrams - non-centered (left) and centered (right) GVA structure data.

Because we were interested in the possible difference between the former East and West Germany, we applied cluster analysis to orthonormal coordinates (see Pawlowsky-Glahn, Egozcue (2005)), defined for a three-part composition $\mathbf{x} = (x_1, x_2, x_2)$ as

$$z_1 = \frac{\sqrt{2}}{\sqrt{3}} \ln \frac{x_1}{\sqrt{x_2 \cdot x_3}}, \; z_2 = \frac{1}{\sqrt{2}} \ln \frac{x_2}{x_3} \tag{1}$$

(up to permutation of parts x_1, x_2, x_3), of GVA structure in federal states and selected big German cities. The conclusion seems to be positive for reunification efforts in Germany because Figure 3 visibly shows that there are no more big differences between the East and the West (in sense of the GVA structure). For example, we can see a cluster of big cities and western states - Saarland, Nordrhein-Westfalen and Hessen on the left side. On the right side there are clusters of mixed eastern and western states. We can also notice here that Mecklenburg-Vorpommern has its own cluster and this is probably caused by fact that this state belongs to the poorest ones.

To see the trend of data we use a regression line and both 95%-confidence band (grey) and simultaneous 95%-confidence band (black) in ternary diagrams and in orthonormal coordinates. The corresponding results are displayed in Figure 4. In the upper right corner there are orthonormal coordinates with x_1 in nominator of z_1 in (1), the lower left corner with x_2 in nominator of z_1 and the lower right corner with x_3 in nominator of the first coordinate, it means that the first coordinate explains relative information on agriculture, production and services, respectively, and the second coordinate represents the remaining log-ratio. We can see that the horizontal (z_1) axis of the first plot of orthonormal coordinates is formed exclusively by negative values; it is caused by the fact that balance consisting of agriculture in nominator is negative - the agriculture part is dominated by the other two components. The trend in data represented by the regression line is almost constant, i.e. there is no systematic trend between the first coordinate and log-ratio of production and services. The other two coordinates show a decreasing trend (the second one is quite steeper) and according to these plots we can see that part of services dominates in the composition.

According to Fišerová and Hron (2012) we tried to form orthogonal regression lines for the analysed composition that could be consequently interpreted as development of each sector of economy in time. Indeed, by interpreting the horizontal line of Figure 5 with parametric form of the regression line in the simplex as time, we can see that in past the highest (and the only) influence on the GVA structure had agriculture ($y_1(t)$), but as the time runs, the influence of agriculture decreases very steeply to zero. At the time of 0, production ($y_2(t)$) and services ($y_3(t)$) curves start to increase (i.e. to take part on relative structure of GVA) significantly. While production increases rather slowly and reaches only point of about 0.3 with the displayed horizontal line limits, services increase steeperly.

From the performed analysis we can see that services mostly partake in relative structure of GVA in German regions; on the other hand, the contribution of agriculture are rather negligible, rounding

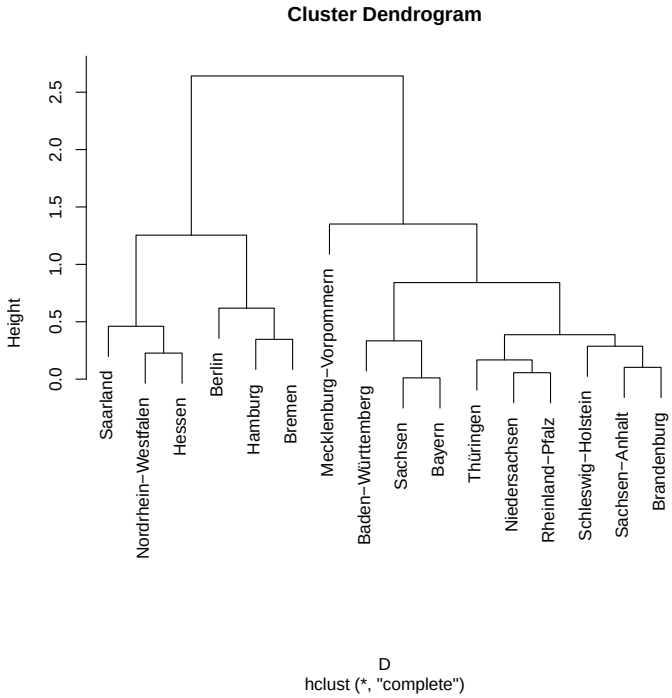


Figure 3: Dendrogram of the GVA structure data.

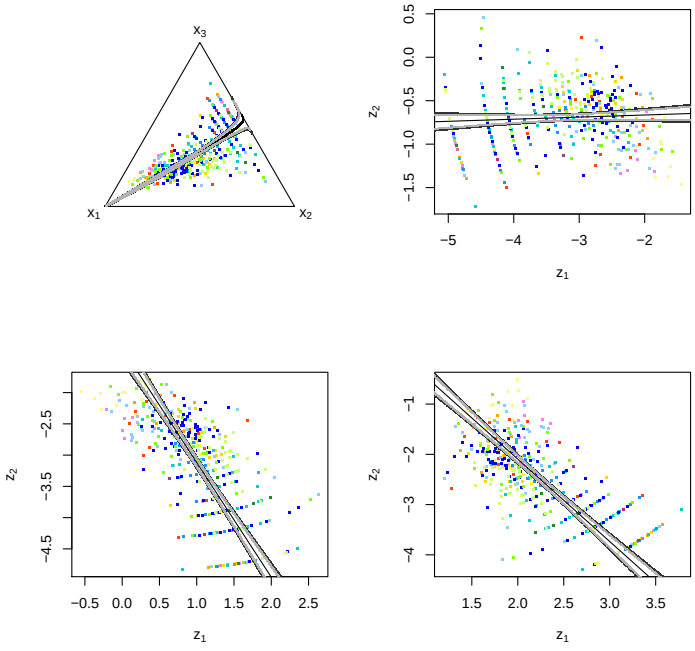


Figure 4: Regression lines of the GVA structure data with the original data in ternary diagram (upper left) and ilr coordinates (upper right, lower left, lower right).

of the small percentage values cause also the line effects in the data structure (both in the ternary diagram and in orthonormal coordinates). Furthermore, we can conclude that results for almost all federal states are very similar (except of Mecklenburg-Vorpommern), i.e. there are recently no more

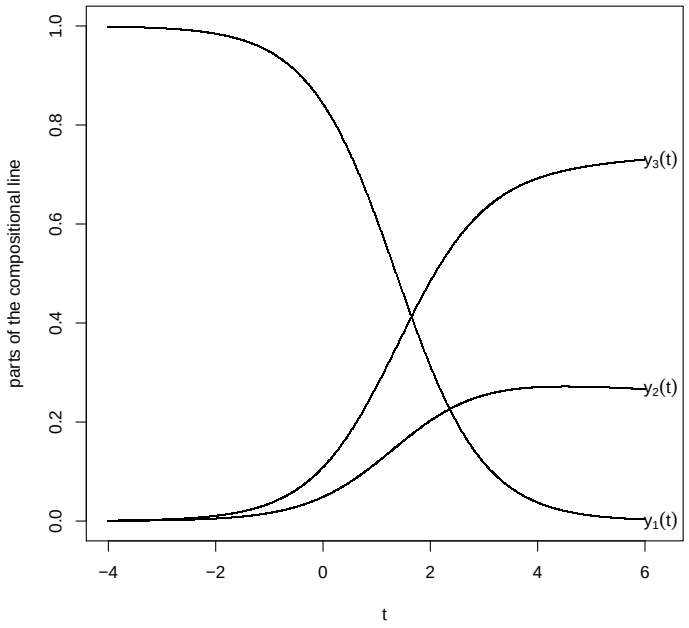


Figure 5: Components of the regression line on the simplex.

big differences between former East and West German regions. However, if we would analyse absolute values of gross value added, the results would be probably very different.

5 Acknowledgement

The authors gratefully acknowledge the support by the Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic).

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.

Fišerová, E. and K. Hron (2012). *Statistical Inference in Orthogonal Regression for Three–Part Compositional Data Using a Linear Model with Type–II. Constraints*. Communications in Statistics - Theory and Methods, 41: 13-14, 2367-2385

Härdle, W. K. and L. Simar (2012). *Applied Multivariate Statistical Analysis*. Springer-Verlag Berlin Heidelberg

Landesbetrieb für Statistik und Kommunikationstechnologie Niedersachsen (2012). *Kreiszahlen: Ausgewählte Regional Daten für Deutschland*. Landesbetrieb für Statistik und Kommunikationstechnologie Niedersachsen, Hannover

Lequiller, F. and D. Blades (2006). *Understanding National Accounts*. OECD [online] http://www.eastafritac.org/images/uploads/documents_storage/Understanding_National_Accounts_-_OECD.pdf

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. and J. J. Egozcue (2005). *Groups of Parts and Their Balances in Compositional Data Analysis*. Mathematical Geology 37, 795-828

Samuelson, P. A. and W. A. Nordhaus (1989). *Economics*. McGraw Hill, Inc., New York

Templ, M., K. Hron, and 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*. John Wiley & Sons, Chichester (UK), pp. 341–355.

Modelling compositional change with simplicial linear ordinary differential equations

E. JARAUTA-BRAGULAT and J. J. EGOZCUE

Department of Applied Mathematics III (DMA3) - School of Civil Engineering (ETSECCPB)
Universitat Politècnica de Catalunya, Barcelona, Spain
eusebi.jarauta@upc.edu

Abstract

In many fields, and particularly in economic and social sciences, compositional data evolving in time or space appear quite frequently. For instance, shares of a market or a portfolio, proportions of social groups from the point of view of the production structure, proportions of GDP by countries or geographic areas, are compositional data changing in time and they can be represented by evolutionary compositions. Their prediction and analysis using suitable models is an important goal.

There is a lack of models in which the different proportions are treated jointly and satisfying the principles of compositional data: scale invariance and subcompositional coherence. Scale invariance requires analyses to be invariant under change of units; subcompositional coherence demands that ratios between parts do not change after reduction of the number of components. First order simplicial linear differential equations satisfy the mentioned requirements and provide flexible enough models for low frequency evolutions. This kind of models can be fitted to data using least squares techniques on coordinates of the simplex. The matrix of the differential equation is interpretable, thus providing a powerful analytical tool.

The evolution of the Spanish population is studied as illustrative example; population is divided into four classes (roughly speaking, children, employees, non-employees, retired) and its evolution in the period 1976-2011 is analyzed. A forecast of these population groups is provided. The fitted model reveals a remarkable instability even using only years previous to the present crisis. Evolution of GDP by geographic areas is also analyzed.

1 Introductory examples

Modelling evolution in time or space of a random vector is a common goal in most sciences. Depending on the framework and the type of sampling, they are given different names: stochastic processes, time series, growth curves, production models, and so on. In most cases, random variables considered are strictly positive; herein, attention is centred in this case. Two examples in social and economic sciences may give an idea of the extent of this topic: population groups in Spain (Example A) and Gross Domestic Product, GDP, by geographic areas in the world (Example B).

Example A: population groups in Spain.

Spanish population is the total number of individuals in Spain which evolves in time; usually data are available by years. In this population, different classification criteria can produce some subgroups. In this study, subgroups are defined by age: age less than 16 years, age between 16 and 65 and, finally, older than 65 years. In the second subgroup, employment situation define other two subgroups: unemployed and the rest. With this classification, Spanish population is divided in four subgroups. Data values in the period 1976–2011 and their corresponding proportions are shown in Figure 1.

Example B: Gross Domestic Product by geographic areas in the world.

World is divided in some geographic areas and total GDP can be computed adding GDP of all the countries belonging to each area. The eight considered areas are: Sub Saharan Africa (SSA), Latin America and Caribbean (LCA), South Asia (SOA), East Asia and Pacific (EAP), Europe and Central Asia (ECA), North America (NAM), Middle East and North Africa (MENA) and the rest of the world (OTH). Figure 2 shows GDP series 1989 - 2011 for each area and corresponding proportions.

K. Hron, P. Filzmoser and M. Templ (eds.)
ISBN: 978-3-200-03103-6

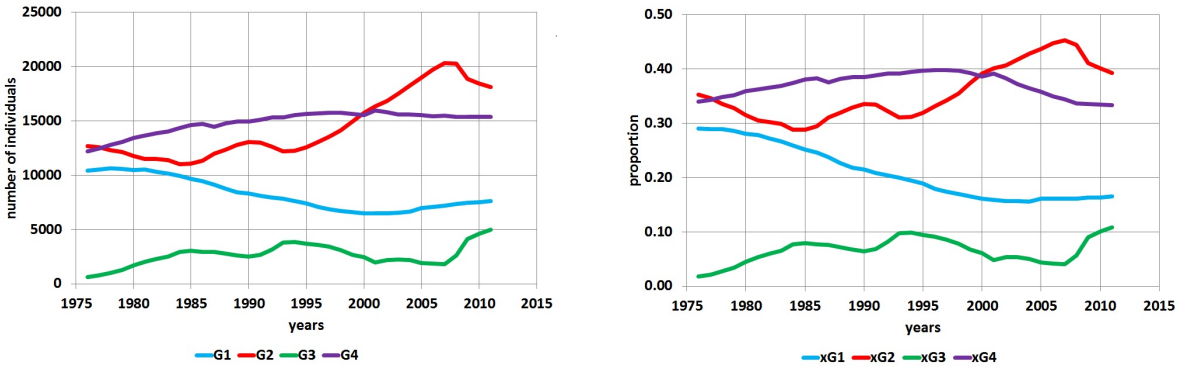


Figure 1: Number of individuals of Spanish population subgroups (left) and proportions (right). G1: age < 16; G2: $16 \leq$ age ≤ 65 and employed; G3: $16 \leq$ age ≤ 65 and unemployed; G4: age > 65. Source of data: Instituto Nacional de Estadística (http://www.ine.es).

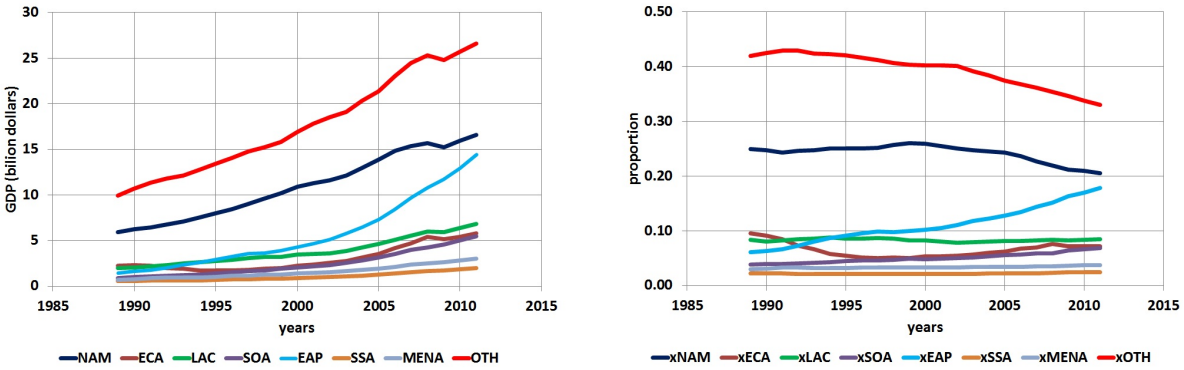


Figure 2: GDP for each area (left) and corresponding proportions (right). Source of data: Google public data (http://www.google.es/publicdata/directory).

In both examples, for each year, a multivariate vector with positive components can be considered; therefore, we have multivariate vectors evolving in time. The evolution of a multivariate vector, with positive components, can be modelled, at least, in two different ways. The first one (traditional modelling) consists in thinking of each component as a real positive function of time. These positive components should frequently be transformed into logarithms due to reasons of their relative scale. A model is then a function $\vec{f} : I \subseteq \mathbb{R} \to \mathbb{R}^m_+$, where I is the parameter interval (time interval, space interval, etc.). A critical point is how the image space $\mathbb{R}^m_+$ is conceived, that is, how distances between two vectors are measured. Frequently, the components of $\vec{f}$ are transformed applying logarithms, so that $\mathbb{R}^m_+$ is transformed into $\mathbb{R}^m$ and the model function is then $\vec{f} : I \subseteq \mathbb{R} \to \mathbb{R}^m$. This means that differences between vectors in $\mathbb{R}^m_+$ are computed as differences of logarithms or log-ratios, thus taking into account the relative character of the components. However, this traditional approach focus the attention in the values of each component, and the ratios between them are in a secondary plane. An alternative proposal consists in paying attention to the ratios between components better than in their values. This is equivalent to consider the data as compositions evolving in time.

The approach proposed herein views the m positive components of the evolving vector decomposed into two different sets of components: a positive component describing a kind of total of the vector, and a compositional vector describing how the total is distributed over the components. For instance, in the Example A (population groups in Spain), total population can be classified into four categories by age and employment. Example B (GDP by geographic areas in the world) considers seven well-defined areas in the world and an additional one as the rest, and it can be studied from the compositional viewpoint.

2 Methodology

Interest is centred in studying functions, from a real interval I into the simplex of m parts, $\mathcal{S}^m$, as models of compositional change in time, space or other real variables. Accordingly, the functions considered are $\vec{f} : I \subseteq \mathbb{R} \to \mathcal{S}^m$. This kind of functions have been called simplex-valued functions (sv-functions) and its elementary calculus has been studied in Egozcue et al. (2011). According to the principle of working in coordinates (Mateu-Figueras et al., 2011), ordinary differential equations for sv-functions can be transformed into orthonormal coordinates (ilr) of the simplex, and then identified and solved using standard available tools of regression and ordinary differential equations in real spaces.

Consider a function $\vec{f} : I \subseteq \mathbb{R} \to \mathbb{R}^m_+$ defined in a real interval I, possibly the whole $\mathbb{R}$. For each $t \in I$, its image $\vec{f}(t) = (f_1(t), \ldots, f_m(t)) \in \mathbb{R}^m_+$ is a vector, where the components verify $f_i(t) > 0$, $i = 1, 2, \ldots, m$, and they represent the evolution of the parts of the modelled system. To support intuition, the components $f_i(t)$ are called *masses* and their sum $M(t)$ is called *total mass*. In general, total mass depends on t but, in some cases, it can be constant. The compositional evolution is described by the closure of the function to a constant, herein assumed to be 1:

$$\mathcal{C}\vec{f}(t) = \frac{1}{M(t)}\vec{f}(t) \quad , \qquad M(t) = \sum_{i=1}^{m} f_i(t) \ . \tag{1}$$

The closed function $\mathcal{C}\vec{f} : I \subseteq \mathbb{R} \to \mathcal{S}^m$ has its values in the simplex of m parts and is a sv-function. Its components are called *parts*. Frequently, the number of parts considered depends on measuring devices, availability of data or easiness of interpretation and, accordingly, the number of parts m is not an objective choice. Then, it is reasonable requiring compatible results, when we consider different number of parts. This corresponds to the subcompositional coherence principle of compositional data analysis. These ideas suggest to use Aitchison geometry of the simplex to analyse the behaviour of a sv-function.

The simplex $\mathcal{S}^m$ is an Euclidean space of dimension $m-1$. Algebraic operations (perturbation $\oplus$ and powering $\odot$) are defined (Aitchison, 1986) as follows: if $\vec{x} = (x_1, x_2, \ldots, x_m), \vec{y} = (y_1, y_2, \ldots, y_m)$ are in $\mathcal{S}^m$ and α is any real number, then

$$\vec{x} \oplus \vec{y} = \mathcal{C}\left(x_1 y_1, x_2 y_2, \ldots, x_m y_m\right), \quad \alpha \odot \vec{x} = \mathcal{C}\left(x_1^\alpha, x_2^\alpha, \ldots, x_m^\alpha\right). \tag{2}$$

These operations conform $\mathcal{S}^m$ as a real vector space. Aitchison geometry is completed with a metric which can be defined using the centered log-ratio (clr) transformation of a composition (Aitchison, 1986; Pawlowsky-Glahn and Egozcue, 2001) as follows:

$$\mathrm{clr}(\vec{x}) = \left(\log\frac{x_1}{g_\mathrm{m}(\vec{x})}, \log\frac{x_2}{g_\mathrm{m}(\vec{x})}, \ldots, \log\frac{x_m}{g_\mathrm{m}(\vec{x})}\right), \quad \sum_{i=1}^{m} \log\frac{x_i}{g_\mathrm{m}(\vec{x})} = 0 \ , \tag{3}$$

where $g_\mathrm{m}(\cdot)$ denotes the geometric mean of the terms in the argument. The inner product and distance in $\mathcal{S}^m$ are defined as

$$\langle \vec{x}, \vec{y} \rangle_a = \langle \mathrm{clr}(\vec{x}), \mathrm{clr}(\vec{y}) \rangle \quad , \quad \mathrm{d}_a\left(\vec{x}, \vec{y}\right) = \mathrm{d}\left(\mathrm{clr}(\vec{x}), \mathrm{clr}(\vec{y})\right), \tag{4}$$

where $\langle \cdot, \cdot \rangle$ and $\mathrm{d}(\cdot, \cdot)$ are the ordinary inner product and distance in a real Euclidean space. With this metric structure, the simplex $\mathcal{S}^m$ is a Euclidean space of dimension $m-1$. Euclidean spaces always admit representation of its elements using coordinates in some orthonormal basis. Isometric log-ratio transformations (ilr) are mappings which assigns coordinates, with respect to an orthonormal basis, to a composition (Egozcue et al., 2003). An ilr transformation is determined by an $(m, m-1)$-matrix V, called *contrast matrix* (Egozcue et al., 2011). Given a contrast matrix V, the corresponding ilr transformation of a composition $\vec{x} \in \mathcal{S}^m$ and its inverse are

$$\vec{u} = \mathrm{ilr}(\vec{x}) = V^\top \log(\vec{x}) \quad , \quad \vec{x} = \mathrm{ilr}^{-1}(\vec{u}) = \mathcal{C}\exp\left(V\vec{u}\right), \tag{5}$$

respectively, where functions log and exp act componentwise ($\top$ indicates transpose matrix).

3 Differential equations in the simplex

With the Aitchison geometry of the simplex in mind, the change of a composition $\vec{x}$ to another $\vec{y}$ is measured by their perturbation-difference, $\vec{x} \ominus \vec{y}$, where $\ominus \equiv \oplus(-1)\odot$, that is, perturbation-subtraction consists of adding the opposite element. Therefore, a natural definition of derivative of a sv-function (Aitchison, 2003; Egozcue et al., 2011), called *simplicial derivative*, is the following: at time t, the simplicial derivative is

$$D^{\oplus}\vec{f}(t) = \lim_{h\to 0}\left(\frac{1}{h}\odot\left(\vec{f}(t+h)\ominus\vec{f}(t)\right)\right), \tag{6}$$

provided the limit exists. Equation (6) matches standard definition of derivative but using the operations in $\mathcal{S}^m$. Computation of simplicial derivatives (Egozcue et al., 2011) can be carried out as

$$D^{\oplus}\vec{f}(t) = \mathcal{C}\exp\left(D\log\left(\vec{f}(t)\right)\right) = \mathcal{C}\exp\left(\frac{Df_1(t)}{f_1(t)}, \frac{Df_2(t)}{f_2(t)}, \ldots, \frac{Df_m(t)}{f_m(t)}\right), \tag{7}$$

where D means ordinary derivative, applied componentwise to vectors.

Two important results concerning simplicial derivative (Egozcue et al., 2011) are the following. The first one states that (7) still applies for any positive-component vector-valued function, that is:

$$D^{\oplus}\mathcal{C}\vec{f}(t) = D^{\oplus}\vec{f}(t) \quad , \quad \vec{f}: I \subseteq \mathbb{R} \to \mathbb{R}^m_+ \; ; \tag{8}$$

the second result relates the simplicial derivative, the ilr-transformation and its ordinary derivative:

$$\mathrm{ilr}\left(\vec{f}(t)\right) = \vec{u}(t) \quad , \quad \mathrm{ilr}\left(D^{\oplus}\vec{f}(t)\right) = D\vec{u}(t). \tag{9}$$

4 Simplicial differential linear models

A simplicial ordinary differential equation is an equation involving the simplicial derivative of a sv-function, the sv-function and, possibly, constants or functions of independent variable t. When equation is linear, it is a *simplicial linear ordinary differential equation* (SLODE). Suppose that a basis in $\mathcal{S}^m$ has been selected and the associated contrast matrix V is given. The composition $\vec{x}(t) \in \mathcal{S}^m$ can be transformed into its coordinates according (5):

$$\vec{u}(t) = V^{\top}\log\left(\vec{x}(t)\right) \quad , \quad \vec{u}(t) \in \mathbb{R}^{m-1}, \tag{10}$$

which can be viewed as a solution of a system of ordinary differential equations. A general linear model is

$$D\vec{u}(t) = A\vec{u}(t) + \vec{b}\,, \tag{11}$$

where $\vec{b} \in \mathbb{R}^{m-1}$ and A is an $(m-1, m-1)$ real matrix of constant coefficients. Equation (11) can be transformed back into the simplex, using ilr^{-1}. Taking into account (5), the corresponding compositional model is

$$D^{\oplus}\vec{x}(t) = [M \boxdot \vec{x}(t)] \oplus \vec{k}, \quad \vec{x}(t), \vec{k} \in \mathcal{S}^m, \tag{12}$$

where $\vec{k} = \mathrm{ilr}^{-1}(\vec{b})$, $M = VAV^{\top}$ is a (m, m) real matrix and the symbol $\boxdot$ indicates the simplicial matrix product (Egozcue et al., 2011).

In practice, coefficients of the model in Eq. (11) are unknown and must be estimated from a dataset. The values of A and $\vec{b}$ determine the behavior of the model. For instance, if A can be considered null, the model in Eq. (12) is simpler than the complete model. The parsimony principle proposes to fit the simplest model, provided that it satisfactorily explains the features of the data. Criteria to decide whether a model is satisfactory or not, should be statistical, e.g. minimization of the norm of residuals, or qualitative.

To select a suitable model, a hierarchy of models with increasing complexity is easily defined from Eq. (11): *Model 0*, in which A is assumed null; *Model 1*, where $\vec{b}$ is assumed null, and *Model 2*,

the complete model . Those proposed models are autonomous, i.e. the variable t does not appear explicitly in the equation, thus assuming that external influences on the system are represented by the constant term $\vec{b}$ (*Models 0 and 2*). Internal relationships between coordinates or parts are expressed by matrices A and M, respectively. External disturbances in the modelled system could be introduced in the model in several ways: a forcing term explicitly depending on time; or a change of the matrices A and M what can be interpreted as an internal change of the system. However, attention is paid only to the three autonomous models mentioned.

Qualitative properties of *Models 0, 1, 2* should be taken also into account when selecting the appropriate model. *Model 0*, with non-neutral $\vec{k}$, is always unstable, i.e. when t tends to infinity some parts go to zero. *Model 1* can be stable or unstable depending on the eigenvalues of A. When all real parts of the eigenvalues are negative, the system is stable and their solutions approach the neutral element of the simplex for increasing time. *Model 2* has more flexibility when it is stable, the solutions can converge to any point in the simplex. These features should be taken into account when making predictions using the fitted model.

5 Model parameter estimation and checking

Residuals are defined as deviations of the data from the values predicted by the model. Estimation of coefficients of the model requires optimizing some target function of residuals. If the data-set is compositional, the natural way of computing deviations is using the perturbation-difference $\ominus$, and the size of the residuals measured by their Aitchison norm. The principle of working on coordinates (Mateu-Figueras et al., 2011) suggests to formulate the problem of model fitting on coordinates of the simplex, thus translating geometric details and computation to the real space of coordinates. Then, a criterion to fit the model can be the least-squares techniques applied to residuals of coordinates. A new approach to estimate the coefficients of the model (11) relies on the integration in time of this differential equation. In fact, integrating (11) from t_1 to any time $t \in I$, it yields,

$$\vec{u}(t) - \vec{u}(t_1) = A\vec{U}(t) + \vec{b}(t - t_1) \quad , \quad \vec{U}(t) = \int_{t_1}^{t} \vec{u}(\tau)\, d\tau \ , \tag{13}$$

where the target parameters A and $\vec{b}$ still appear as linear parameters of the model. We use regression to estimate A and $\vec{b}$. The model (13) appears as a multivariate regression problem once the observations, $\vec{u}_o(t_j)$ are inserted in it:

$$\vec{u}_o(t_j) = \vec{u}(t_1) + A\vec{U}(t_j) + \vec{b}(t_j - t_1) + \vec{e}_j, \quad j = 1, 2, \ldots, n \ , \tag{14}$$

where A and $\vec{b}$ are the matrix of regression coefficients, and the components of $\vec{u}(t_1)$ are the intercepts. The residuals of the regression model are $\vec{e}_j$. Numerical integration to estimate values of integral functions are made applying trapezoidal rule. The statistical analysis is performed under bootstrap resampling.

Integrated model (13) has been selected due to, at least, two reasons. The first one is that the primary interest is in the coordinates $\vec{u}(t)$, and therefore, residuals are measured as differences on them. Alternatively, using (11), residuals would be measured as differences of derivatives. The second reason is that using (13) does not require estimation of derivatives; they are replaced by a numerical estimation of an integral function. Normally, numerical integrals are more accurate than numerical derivatives. Accuracy and stability of exposed models can be illustrated showing bootstrap credible intervals and median for the evolutionary compositions. This graphical representation is a useful tool for deciding which is the most appropriate model to use in each case.

6 Applications

Different aspects and details of the exposed approach are commented in this section developing the two examples introduced in Section 1.

Table 1: Example A. Multiple regression coefficients R^2 obtained for each model and the three balances. The number of parameters fitted in each regression is shown under *param*.

	u_1	u_2	u_3	param.
Model 0	0.01572	0.77539	0.77381	3
Model 1	0.70947	0.99599	0.92217	9
Model 2	0.84010	0.99652	0.95818	12

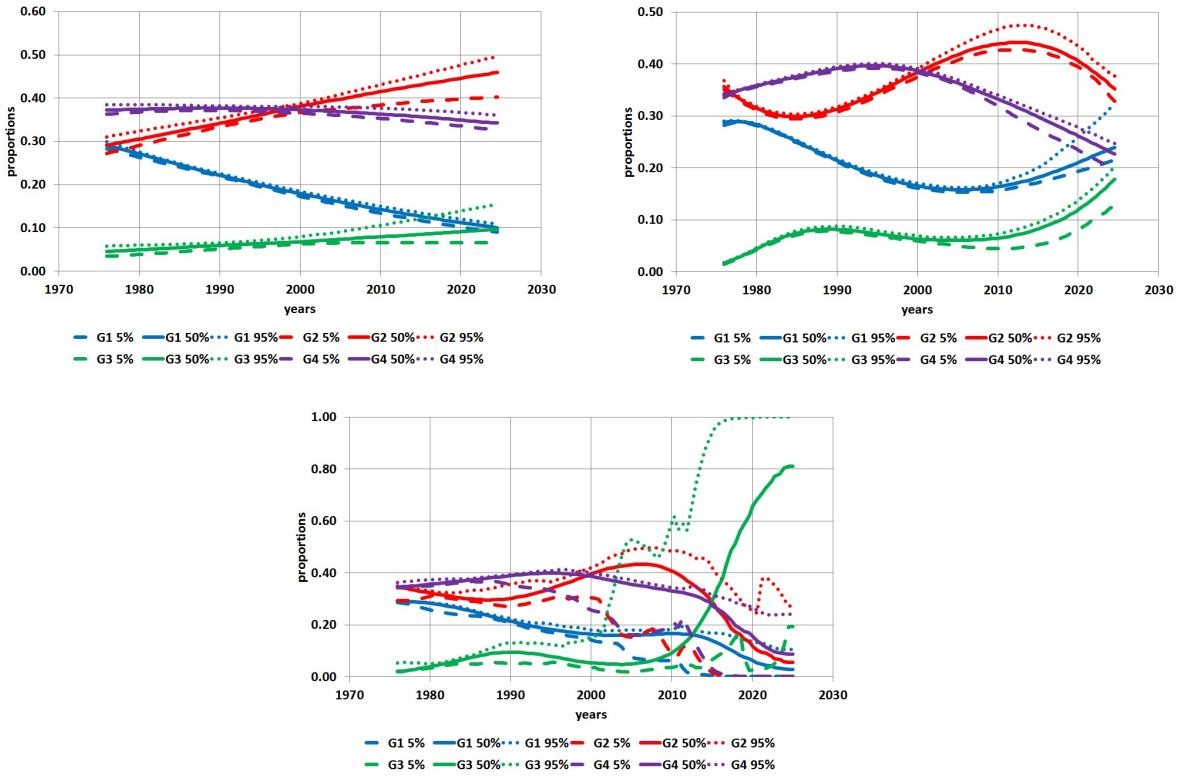


Figure 3: Example A. 90% bootstrap credible intervals (dots, upper 95% quantile; continuous line, median; dashes, lower 5% quantile) corresponding to *Model 0* (left), *Model 1* (right) and *Model 2* (down).

6.1 Example A: compositional evolution of population groups in Spain.

The Spanish population, 1976−2011, has been classified into four groups as explained in Section 1. Figure 1 shows the absolute population in each category (left panel) and the corresponding proportions (right panel). The contrast matrix used to obtain ilr-coordinates is:

$$V = \begin{pmatrix} 0.00000 & 0.70711 & -0.50000 \\ 0.70711 & 0.00000 & 0.50000 \\ -0.70711 & 0.00000 & 0.50000 \\ 0.00000 & -0.70711 & -0.50000 \end{pmatrix}.$$

In Table 1, multiple regression coefficients R^2 obtained for the three models and the three balances are shown. The number of yearly data points is $n = 36$. Figure 3 shows bootstrap credible intervals and median for the evolutionary compositions; this figure illustrates also the effect of overparametrization produced applying *Model 2*. According the procedure proposed, *Model 1* is considered the most appropriate one. In Figure 4, time evolution of balances and proportions is shown for *Model 0* and *Model 1*.

Comments related to this example (see Figures 3, 4):

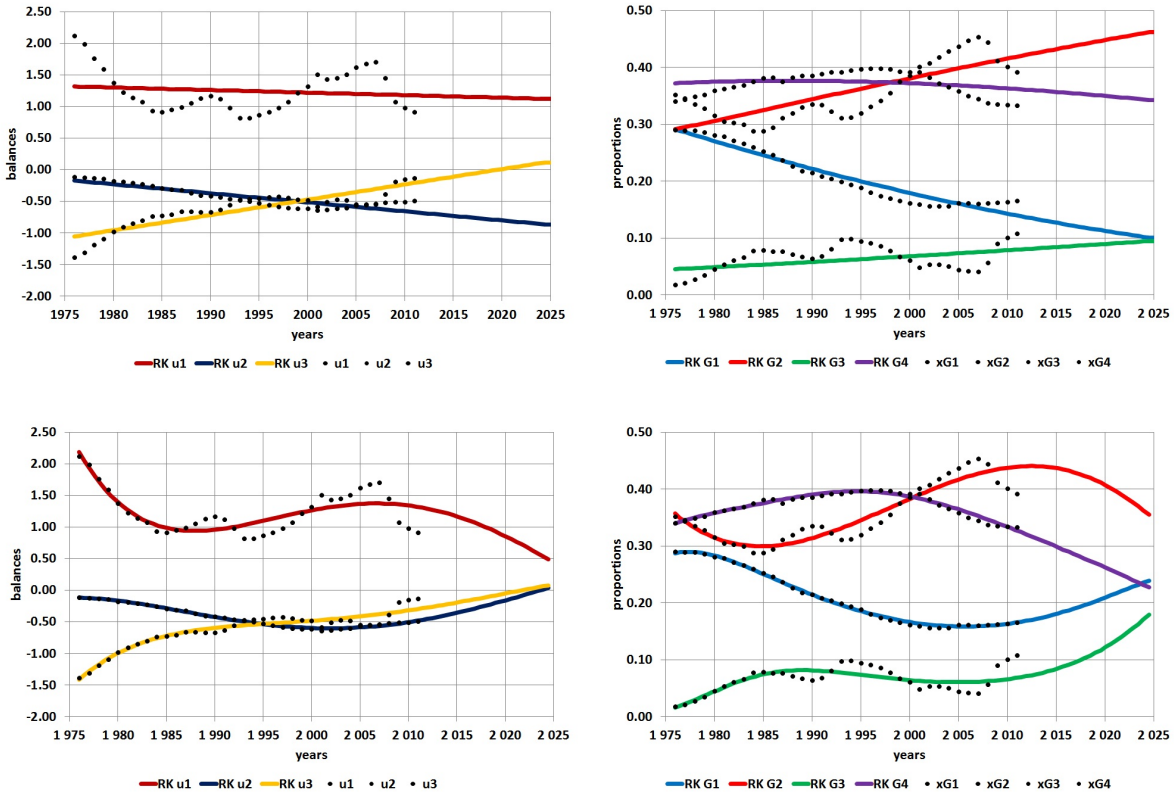


Figure 4: Example A. Time evolution of balances (left) and proportions (right) for *Model 0* (first row) and *Model 1* (second row). Dots in black are data balances (left) and data values (right). RK indicates Runge-Kutta solutions.

Table 2: Example B. Multiple regression coefficients R^2 obtained for each model and the seven balances. The number of parameters fitted in each regression is shown under *param*.

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	param.
Model 0	0.10352	0.88754	0.94339	0.94479	0.92794	0.27326	0.87529	3
Model 1	0.99319	0.98865	0.99307	0.99731	0.97195	0.95002	0.99878	9
Model 2	0.99329	0.98904	0.99346	0.99918	0.98121	0.95002	0.99890	12

- The compositional model applied for the analysis, takes into account better the interrelationships between population groups than the analysis of single components based on the evolution of the number of individuals of the population groups.
- The criterion used for the classification of the population allows distinguishing groups with demographic criteria (G1 and G4) and labor market criteria (G2 and G3).
- Data related to groups G1 and G4 are only affected by demographics. The estimated model fits better the behavior of G1, G4 than the groups G2 and G3, which are affected by fluctuations in labor market.
- Tendency to decrease from year 2011 in the proportion of the group G2 (employees) and trend towards increase of the group G3 (non-employees), from year 2010 onwards, is also obtained when using the data set previous to 2007, before the onset of the global economic crisis.
- The trend towards the increasing proportion of the youngest population (group G1) is interpreted by the influence of immigration in Spain along the past decade.
- The decrease in the proportion of older (G4 group) population can be due to several reasons. However, it suggests a decrease in life expectancy, but it can not be decided from a purely compositional analysis.
- The estimated SLODE is unstable. This means that there should be a time for which the model cannot reflect the dynamics of the population, thus suggesting a change of paradigm.
- *Model 2* is more flexible and predicts similar features to those given by *Model 1*. However, as shown in the bootstrap credible intervals, *Model 2* is over-parametrised (Figure 3, bottom).

6.2 Example B: compositional evolution of GDP of world areas.

The classification of world economic areas in eight groups has been introduced in Section 1. Data values in the period 1989–2011 and its corresponding proportions are shown in Figure 2. The contrast matrix used to obtain ilr-coordinates in this case is:

$$V = \begin{pmatrix} 0.70711 & 0.40825 & 0.28868 & 0.22361 & 0.18257 & 0.15430 & 0.13363 \\ -0.70711 & 0.40825 & 0.28868 & 0.22361 & 0.18257 & 0.15430 & 0.13363 \\ 0.00000 & -0.81650 & 0.28868 & 0.22361 & 0.18257 & 0.15430 & 0.13363 \\ 0.00000 & 0.00000 & -0.86604 & 0.22361 & 0.18257 & 0.15430 & 0.13363 \\ 0.00000 & 0.00000 & 0.00000 & -0.89444 & 0.18257 & 0.15430 & 0.13363 \\ 0.00000 & 0.00000 & 0.00000 & 0.00000 & -0.91285 & 0.15430 & 0.13363 \\ 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & -0.92580 & 0.13363 \\ 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & 0.00000 & -0.93541 \end{pmatrix}.$$

In Table 2, multiple regression coefficients R^2 obtained for the three models and the seven balances are shown. The number of yearly data points is $n = 23$. Figure 5 shows bootstrap credible intervals and median for the evolutionary compositions; Figure 5 also illustrates the effect of overparametrisation produced applying *Model 1* and *Model 2*. *Model 0* is considered as the most appropriate since symptoms of over-parametrisation are quite evident in Figure 5. In Figure 6, time evolution of balances and proportions obtained applying *Model 0* is shown.

Comments related to this example (see Figures 5 and 6):

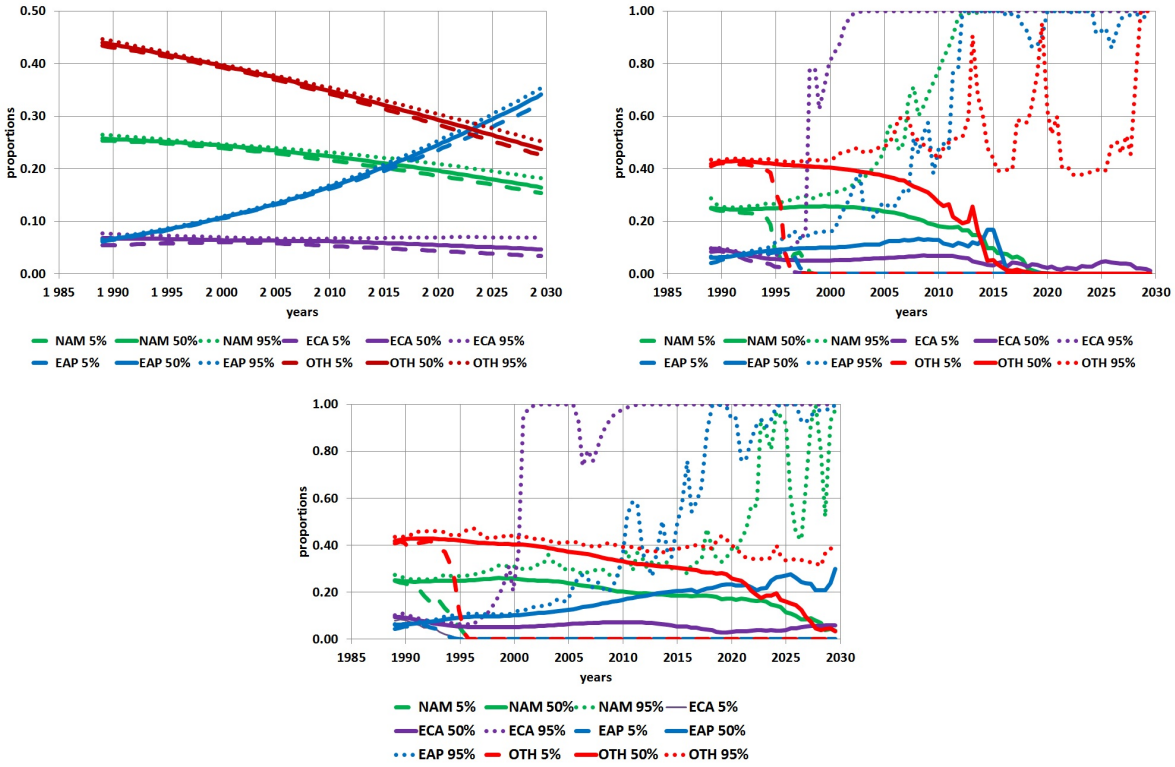


Figure 5: Example B. 90% bootstrap credible intervals (dots, upper 95% quantile; continuous line, median; dashes, lower 5% quantile) corresponding to *Model 0* (left), *Model 1* (right) and *Model 2* (down). Only four areas (NAM, ECA, EAP, OTH) are shown to make picture interpretable and illustrative.

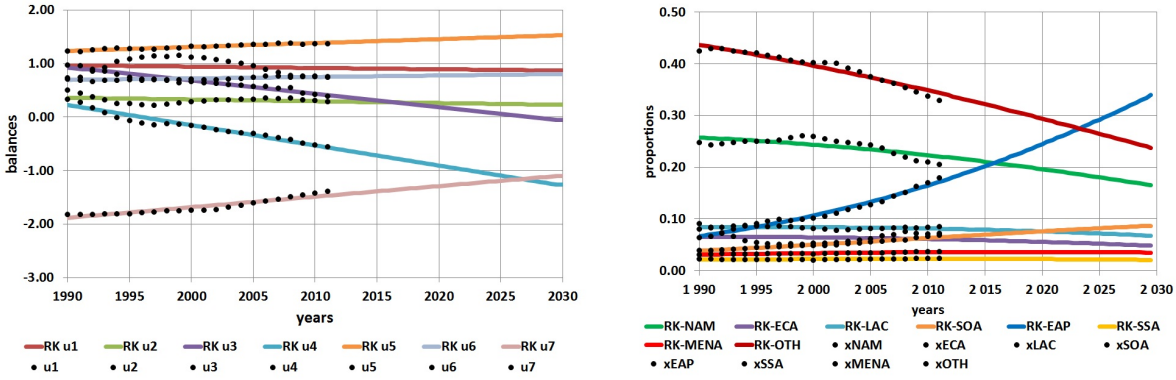


Figure 6: Example B. Time evolution of balances (left) and proportions (right) for *Model 0*. Dots in black are data balances (left) and data values (right). RK indicates Runge-Kutta solutions.

- The analysis, takes into account better the interrelationships between world areas than single component analysis based on the evolution of the GDP of each area.
- The evolutionary trend of the proportions shows surprising values in some cases and disturbing in others. They should motivate the study of measures to correct the negative effects that can be observed.
- A decreasing trend of the weight of the ECA (Europe and Central Asia) area in the global context is shown.

7 Conclusions

Additionally to comments and specific conclusions for Examples A and B, some more general conclusions can be stated:

- Compositions evolving in time or space are common in most fields of science. However, there is a lack of models for these compositional phenomena.
- The advances in compositional data analysis provided new analytical tools such as the simplicial differential equations and their treatment using isometric log-ratio coordinates.
- Simplicial differential equations in their simplest versions, i.e. first order, constant coefficients, autonomous, linear differential equations, are rich evolutionary models in many situations.
- A new fitting technique, based on integration of the model and regression techniques, is proposed.
- Bootstrap techniques have been shown useful for model validation.

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.

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

Egozcue, J. J., C. Barceló-Vidal, J. A. Martín-Fernández, E. Jarauta-Bragulat, J. L. Díaz-Barrero, and G. Mateu-Figueras (2011). *Elements of simplicial linear algebra and geometry*. In: Pawlowsky-Glahn, V. and Buccianti A. (eds.), *Compositional Data Analysis: Theory and Applications*, Wiley, Chichester UK.

Egozcue, J. J., E. Jarauta-Bragulat, and J. L. Díaz-Barrero (2011). *Calculus of simplex-valued functions*. In: Pawlowsky-Glahn, V. and Buccianti A. (eds.), *Compositional Data Analysis: Theory and Applications*, Wiley, Chichester UK.

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.

Mateu-Figueras, G., V. Pawlowsky-Glahn, and J. J. Egozcue (2011). *The principle of working on coordinates*. In: Pawlowsky-Glahn, V. and Buccianti A. (eds.), *Compositional Data Analysis: Theory and Applications*, Wiley, Chichester UK.

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.

Partial least squares for compositional data used in metabolomics

A. KALIVODOVÁ[12], K. HRON[12], M. ŽUPKOVÁ[3], H. JANEČKOVÁ[3] and D. FRIEDECKÝ[3]

[1]Department of Mathematical Analysis and Applications of Mathematics - Palacký University, Czech Republic
Kalivodovaa@gmail.com
[2]Department of Geoinformatics - Palacký University, Czech Republic
[3]Laboratory of Metabolomics - Institute of Molecular and Translational Medicine, Palacký University, Czech Republic

1 Metabolomics

The metabolome is a collection of small molecular mass organic compounds which are in a given biological material. It includes all organic substances naturally occurring from the metabolism of the studied living organism. Molecules that form the metabolome are called metabolites. The analysis of metabolome in a given condition is called metabolomics. Metabolomics is related with a data analysis, it is based on the interpretation of information-rich data aimed at complementing the understanding of biological processes (Roux et al., 2011).

Metabolomics has several strategic approaches for analysis of the biological sample. One of them is the targeted analysis of the metabolite. This approach is focused on the analysis of the specific group of metabolites related to certain metabolic pathway or a class of compounds. Therefore, some metabolic information is usually ignored. A sample preparation is very complex and difficult. Another approach is more complex and it is called untargeted metabolomics. The untargeted approach is a global analysis of metabolic changes in response to disease, environmental or genetic deviation. The untargeted approach is typically carried out for a hypothesis of generation, followed by targeted profiling for a more confident quantitation of relevant metabolites (Xiao et al., 2012).

Data in metabolomics have a specific structure. Metabolites are measured on some biological material (for example cells or blood) and they carry relative information. Concerning the statistical analysis, the problem occurs that more metabolites (in hundreds) than biological materials (only tens) are present in these data sets. Therefore, suitable methods must be used for these data. One of them is the partial least squares regression (PLS regression) that additionally needs to be adapted for the compositional case.

2 Compositional data

When quantifying an information in metabolomics, the results are often expressed in the form of data carrying only relative information. Vectors of these data have positive components and the only relevant information is contained in the ratios between their parts, i.e. we are forced to deal with compositional data (Aitchison, 1986). In such a case, as usual, the sum of the variables (parts) can be rescaled to a prescribed constant.

The standard preprocessing of the compositional data is following. The zero replacement of zeros under the detection limit is made first and then any "reasonable" transformation is applied. It is common to use a logarithmic transformation in chemistry and medicine. This is not used in our approach. Because metabolomical data have properties of compositional data, logratio transformations are used. The best way is to use isometric logratio (ilr) transformation. The principle of this transformation is to form an orthonormal basis for a composition $\mathbf{x} = (x_1, \ldots, x_D)'$ on the simplex (sample space). There are several possibilities of constructing such a basis. One of them results in $(D-1)$-dimensional real vector $\mathbf{z} = (z_1, \ldots, z_{D-1})'$ (Hron et al., 2012; Egozcue and Pawlowsky-Glahn, 2005), which components are defined as

$$z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j}}, \quad i = 1, \ldots, D-1. \tag{1}$$

The variable z_1 carries all the relevant information about the compositional part x_1, i.e., it explains all the ratios between x_1 and the other parts of $\mathbf{x}$ (Hron et al., 2010, 2012). If the parts $x_2, \ldots, x_D$ are permuted, the interpretation of z_1 remains unchanged.

Now we can permute the indices in formula (1), because we want to construct an orthonormal basis, where the first ilr coordinate explains the compositional part of interest. The part of interest, x_l, $l = 1, \ldots, D$, plays the role of x_1 there. It is thus necessary to construct D different ilr transformations, where the D-tuple $(x_1, \ldots, x_D)$ in (1) is replaced, e.g. by $(x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D) =: (x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})$ (Hron et al., 2012). Consequently, the ilr transformation results in

$$z_i^{(l)} = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}}, \quad i = 1, \ldots, D-1. \tag{2}$$

The above form of the ilr transformation is convenient for a regression analysis. For the purpose to construct the biplot for better visualization of data in metabolomics, the centered logratio (clr) transformation seems to be more advantageous. Clr transformation results in a real vector

$$\mathbf{y} = (y_1, \ldots, y_D)' = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right)'. \tag{3}$$

It is possible to express a linear relation between the clr coefficients and the ilr coordinates as $\mathbf{y} = \mathbf{V}\mathbf{z}$. The matrix $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_{D-1})$ has dimension $D \times (D-1)$ and its columns are formed by the clr transformed orthonormal basis vectors (with respect to the Aitchison geometry) (Filzmoser et al., 2012),

$$\mathbf{v}_{D-i} = \sqrt{\frac{i}{i+1}} \left(0, \ldots, 0, 1, -\frac{1}{i}, \ldots, -\frac{1}{i} \right)', \quad i = 1, \ldots, D-1. \tag{4}$$

Approaches for a metabolomic data analysis using logarithmic and logratio transformations will be compared in the next sections.

3 Partial least squares regression

3.1 Standard approach

The Partial Least Squares (PLS) is a class of methods for a modeling of relations between sets of observed variables by means of latent variables (Rosipal and Kramer, 2006). This method can be considered as a combination of principal component analysis and a multiple regression. The PLS is the most widely used method in chemometrics for the multivariate calibration even though the fact that it was first used in social studies (Varmuza and Filzmoser, 2009). The aim of the PLS is to predict or analyze a set of dependent variables from a set of independent variables or predictors. The prediction is made by extracting a set of orthogonal factors from the predictors which are called latent variables. The goal is to maximize the covariance between different sets of variables (Rosipal and Kramer, 2006).

The PLS regression is a technique to relate the data matrix $\mathbf{X}$ to the vector $\mathbf{y}$ or to the matrix $\mathbf{Y}$. Matrices $\mathbf{X}$ and $\mathbf{Y}$, respectively the vector $\mathbf{y}$, are considered as blocks of variables. Two basic PLS approaches are called PLS1 and PLS2. First of them works with only one response variable, second has more response variables. Therefore, in case of the PLS1 only the vector $\mathbf{y}$ is used instead of the matrix $\mathbf{Y}$ as in PLS2 (Rosipal and Kramer, 2006).

The PLS2 regression is more useful in case of chemometric data. We can examine more groups of patients and controls with specific diseases in this approach. Let us consider multivariate $x-$ and $y-$data given by the matrix $\mathbf{X}$ of the dimension $n \times D$ and the matrix $\mathbf{Y}$ of the size $n \times q$, respectively. Data in rows in the matrix $\mathbf{X}$ represent n objects with D features (predictor variables), $\mathbf{Y}$ describes q properties (response variables). So we have q groups of controls and patients with some

disease. x-data are transformed into a set of few latent variables, scores, and these new variables are used for the regression with dependent variables $\mathbf{Y}$. The criterion for the latent variables is the maximum covariance between $x-$scores and $y-$scores. This covariance includes the high variance of $\mathbf{X}$ (responsible for the stability) and the high correlation with the interesting property (Varmuza and Filzmoser, 2009). The aim of the PLS2 regression is to find a linear relation between $x-$ and $y-$variables, using the $D \times q$ matrix $\mathbf{B}$ of regression coefficients, and the error matrix $\mathbf{E}$ (Mevik and Wehrens, 2007; Varmuza and Filzmoser, 2009)

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}. \tag{5}$$

Consequently, the objective of PLS1 is to maximize the covariance between x and y-scores.

3.2 Compositional approach

The procedure of the PLS regression is slightly different for compositional data. Matrices $\mathbf{X}$ and $\mathbf{Y}$ have the same structure as before, but now they are matrices of compositions. Both $\mathbf{X}$ and $\mathbf{Y}$ are mean centered (the latter one with respect to the Aitchison geometry). Consequently, the matrix $\mathbf{X}$ is transformed into $\mathbf{Z}$ using the ilr transformation (1). The PLS regression problem has now the multivariate form of the linear regression,

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{6}$$

where $\boldsymbol{\gamma}$ is the $(D-1) \times q$ matrix of regression coefficients.

With the above interpretation, the ilr transformation (1) allows only the interpretation of elements in the first row of $\boldsymbol{\gamma}$. Only the first column of $\mathbf{Z}$ can be assigned with the first compositional part. For associations to other parts we have to use a permutation of these parts. This leads to the ilr transformation of the form (2), where variables $\mathrm{z}_1^{(l)}$, $l = 1, \ldots, D$ describe all the relative information about the first permuted compositional part. The corresponding first row of the matrix $\boldsymbol{\gamma}$ of regression parameters has a similar interpretation.

The outlined procedure suggests to employ PLS regression D times, so that each compositional part occurs once at the first position in the permuted version of the ilr transformation. In addition, there is an orthogonal relation between different ilr transformations (Egozcue et al., 2011). This implies that regression coefficients need to be estimated just once, and all the other solutions can be derived using orthonormal transformations (the hereat permutation of elements of the matrix $\mathbf{V}$ is used to construct such transformations) of the initial estimated regression coefficients.

4 Practical application

The theory from sections above is now applied to a real-world problem from the Laboratory of Metabolomics from the Institute of Molecular and Translactional Medicine.

Data used for the real example were obtained by targeted metabolomic analysis of plasma samples for the diagnosis of inherited metabolic disorders. Plasma samples were analyzed with the flow injection analysis method. All the experiments were performed on a QTRAP 5500 tandem mass spectrometer (AB SCIEX, U.S.A.) with electrospray ionization. The compounds were measured in a multiple reaction monitoring mode. There were analyzed 50 control samples and 27 samples with defects in amino acid metabolism, organic acidurias and mitochondrial defects. From patient samples, one common group was formed that leads to PLS1 (we use only the vector $\mathbf{y}$ instead of the matrix $\mathbf{Y}$ in (5), respectively (6)).

The data set used in this example has 77 rows (patients and controls) and 162 columns (metabolites). Therefore, there are more columns then rows and we have to take this fact into an account.

The first statistical tool, where results of different approaches to the statistical analysis of the data set are compared, is a biplot of a principal component analysis. It is popularly used by chemists and doctors for displaying data in order to see the multivariate data structure and relations between variables. There are many possibilities how to calculate scores and loadings and display the biplot. All results are shown in the Figure 1. The first alternative is to use the original data set and do only

scaling and centering. The second possibility is to apply a standard logarithm to data, a very popular approach in medicine. Metabolomical data have features of compositional data as written above, thus the compositional biplot is used as well. The clr transformation is applied here, because we want to see all variables on the biplot (that would not be the case with the ilr transformation). Consequently, also the PLS regression is applied to clr transformed data. Scores and loadings for a biplot construction are provided by the PLS regression method itself. Finally, a standard logarithm is used prior to the PLS regression is employed for computing scores and loadings of the biplot. Because of a large amount of variables, only 35 most significant metabolites (based on lengths of the loading vectors) are displayed in the biplot.

We don't see any clusters of patient samples in the graph (a) in the Figure 1. Controls are spaced in the whole area of the graph. All arrows have the same direction. This approach thus seems to be completely useless. In (b), some clusters of patients are visible (for example the disease 4 denoted ♦). Arrows have one common direction again; there is only one exception - the metabolite C3, with direction to samples denoted ♦ and ×. C3 is one of markers for patients with these diseases. The graph (c) looks better than the two previous. Clusters of patients are more visible and arrows show more directions. The metabolite C3 has the longest arrow, thus this variable has a large influence on the multivariate data structure, shown in the biplot. Another interesting arrow represents the metabolite denoted as Val. It goes to the direction of the cluster of patients denoted as ■. This metabolite is again one of markers of this corresponding disease. In (d), controls are separated from patients. We can see clusters of patients again. Metabolites C10 or C5.1 distinguish patients from controls. In the last graph (e), patients denoted ♦ and × moved to the cluster of the other patients. This is not good, because they differ from others. If we compare these graphs, we conclude that the (d) represents the optimal way how to construct a biplot. Thus, the best option here is to use a compositional approach to PLS regression with the clr transformation.

In the previous part we concluded that the PLS regression is a good tool to display data in a biplot as well. But here, the presented approach to the PLS regression with the ilr transformation has not been applied. This is used now, when a comparison of a significance of the standardized regression coefficients is perfomed. In case of the PLS regression, the cross-validation was used for this purpose and results of different approaches are collected in the Figure 2.

The first graph denoted (a) in the Figure 2 shows a standard approach tov the estimation of standardized regression coefficients. Data are only scaled there. We can see that in the case of 162 variables, most of corresponding regression coefficients are marked as significant (they are located above/below the cut-off line, represented by the 0,975-quantile of the standard normal distribution). In the graph (b) log-transformed data are used. The result is similar as in the case (a). The graph (c) shows again a similar outcome (although different coefficients from previous cases were marked here), thus a metabolomic interpretation would be necessary to justify the reasonability of the compositional approach also from this point of view. In the last graph (d), results of t-tests, applied to single variables, are displayed. The t-test is a popular statistical tool even for multivariate data in medical applications. It is a one-dimensional statistical tool and it is used for the analysis of multivariate data. However, although we can see less significant points, that could even have some interpretations, this outcome is obvously incorrect and should be avoided.

5 Conclusion

The compositional approach to the analysis of data from metabolomics was introduced. The PLS regression in standard and compositional representations were discussed. Several approaches to the biplot construction were compared. In this comparison, we can conclude that the best approach is to use the clr transformation and then to apply the PLS regression. Results obtained from this graph are meaningful. The comparison of significances of the standardized regression coefficients in the PLS regression (for ilr transformed compositional data, as proposed in the Section 3.2) for different approaches was performed at the end of this paper. Here a more detailed insight into metabolomical aspects of data would be necessary to evaluate the performance of employed approaches.

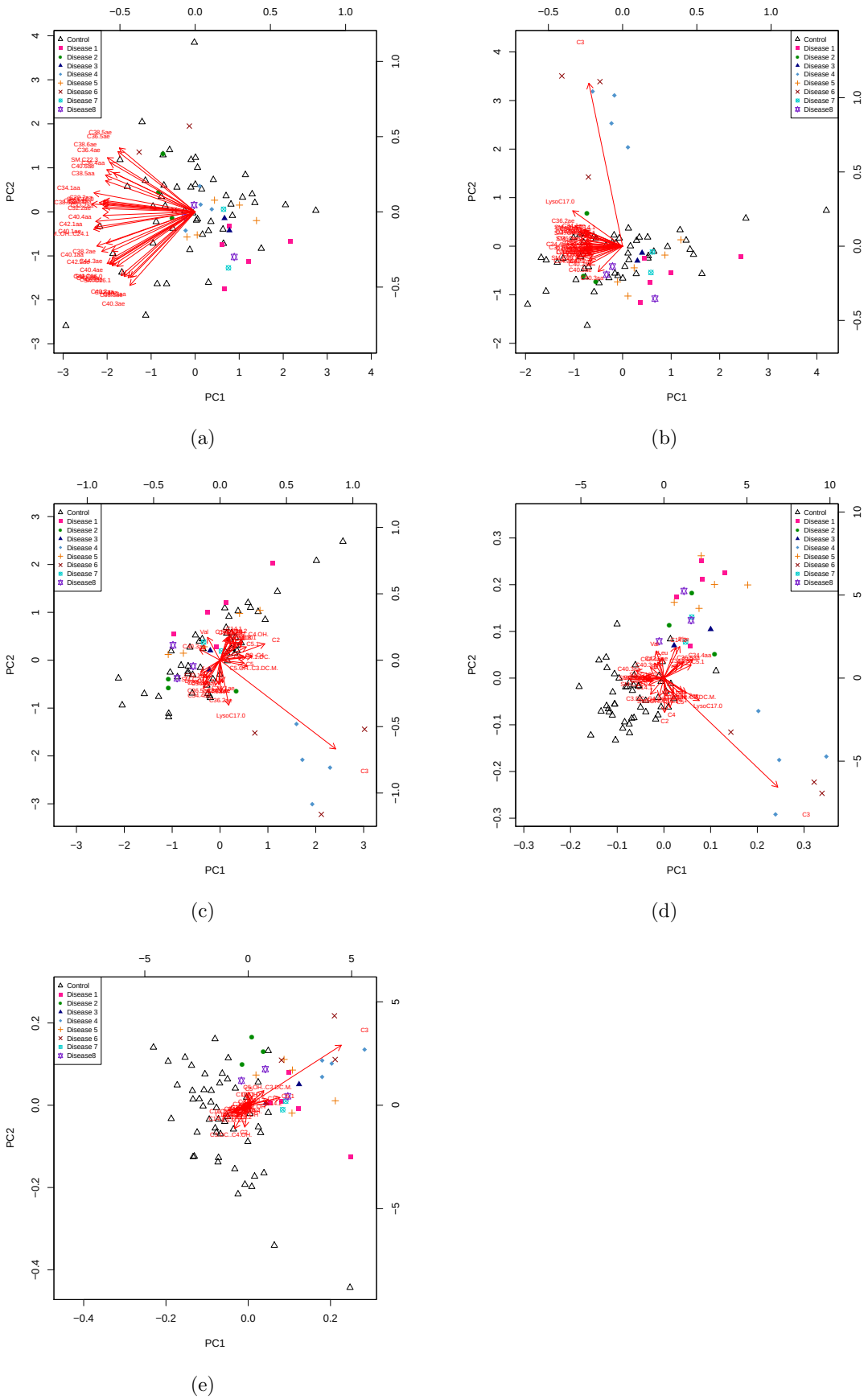


Figure 1: Biplots used for variously preprocessed data sets: (a) original data set with scaling, (b) standard logarithm, (c) compositional attitude with clr transformation, (d) PLS regression on compositional data with clr transformation, (e) PLS regression on data with standard logarithm.

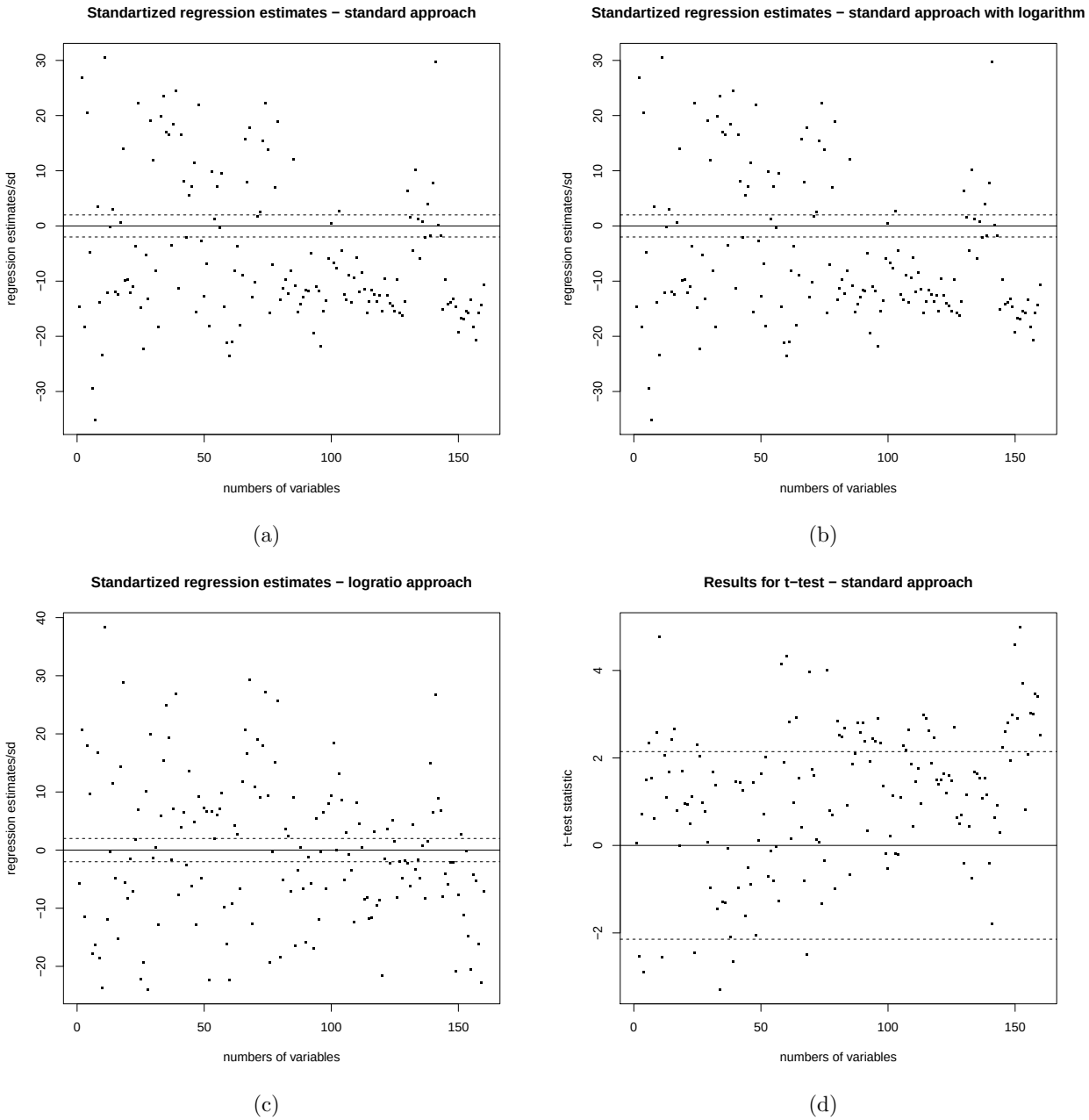


Figure 2: Significances of standartized regression coefficients in the PLS regression for: (a) original data set with scaling, (b) standard logarithm, (c) compositional attitude with ilr transformation, (d) t-test.

6 Acknowledgement

The authors gratefully acknowledge the support by the Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic). The infrastructural part of this project (Institute of Molecular and Translational Medicine) was supported by the Operational programme Research and Development for Innovations (project CZ.1.05/2.1.00/01.0030). The project was supported by grant LF UP 2013-010.

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.

Hron K., P. Filzmoser and K. Thompson (2012). Linear regression with compositional explanatory variables. *Journal of Applied Statistics* 39 (5), 1115–1128.

Hron K., M. templ and P. Filzmoser (2010). Imputation of missing values for compositional data using classical and robust methods. *Comput. Stat. Data Anal.* 54, 3095–3107.

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., C. Barceló-Vidal, J. Martín-Fernández, E. Jarauta-Bragulat, J. Díaz-Barrero, and G. Mateu-Figueras, (2011). *Elements of simplicial linear algebra and geometry*. Compositional data analysis: Theory and applications. Wiley, Chichester, 139–145.

Filzmoser, P., K. Hron and C. Reimann, (2012). Interpretation of multivariate outliers for compositional data. *Computers & Geosciences* 39, 77–85.

Mevik, B.-H. and R. Wehrens (2007). The pls Package: Principal Component and Partial Least Squares Regression in R. *Journal of Statistical Software* 18 (i02), 1–24.

Rosipal, R. and N. Kramer (2006). *Overview and Recent Advances in Partial Least Squares*. SLSFS, Springer, 34–51.

Roux, A., D. Lison, Ch. Junot, and J. Heilier (2011). Chemometrics in Metabonomics. *J. Proteome Res* 6 , 469–479.

Trygg, J., E. Holmes and T. Lundstedt (2007). Applications of liquid chromatography coupled to mass spectrometry-based metabolomics in clinical chemistry and toxicology: A review. *Clinical Biochemistry* 44 , 119–135.

Varmuza, K. and P. Filzmoser (2009). *Introduction to Multivariate Statistical Analysis in Chemometrics*. Taylor & Francis, New York. 336 p.

Xiao J. F., B. Zhou and H. W. Ressom (2012). Metabolite identification and quantitation in LC-MS/MS-based metabolomics. *Trends in Analytical Chemistry* 32, 1–14.

Performance analysis of wastewater treatment in constructed wetlands

J. LIN-YE[1]

[1]Graduate student - UPC Barcelona Tech, Spain jl.iccp@gmail.com

Abstract

The use of horizontal subsurface flow constructed wetlands (CW SSF) for the treatment of municipal wastewater in small communities is becoming a commendable alternative to conventional wastewater management in a society led to the sustainable living. From 2001 to 2003, the Environmental Engineering Division of the Hydraulics Department of the UPC Barcelona Tech had carried out an experiment on CW SSFs in the small town of Les Franqueses, Spain. The CW SSFs were connected in parallel to a wastewater source so that each CW SSF received the same discharge along the observation time. The CW SSFs were designed with different aspect ratios (ratio "length : width" of each CW SSF): A(1:1), B(1.5:1), C(2:1) and D(2.5:1). In type A CW SSFs, water depth was 0.5 m; in type B CW SSFs it was 0.5 m in 2001 and 2002 and changed to 0.27 m in 2003; in type C CW SSFs, it was 0.5 m and for CW SSFs D it was 0.27 m. For each of the CW SSF groups described, there were 2 subgroups with different substrate medium sizes (D_{60}) and porosities (ϕ): subgroup 1 had a substrate medium size of 10 mm and a porosity of 39% while subgroup 2 had a substrate medium size of 3.5 mm and a porosity of 40%. Additionally, a variety of hydraulic load rates (HLR) were applied in order to test HLR effects on CW SSF water depuration: 20, 27, 36 and 45 mm/day. Concentrations of affluent and effluent flow were taken in 824 measurements. This data was analyzed by J. García et al. (2005) using conventional statistics.

Compositional data analysis and linear models are used in the present analysis, which is meant to provide complementary information to the analyses carried out by García. The results determine that the most efficient CW SSFs have the following features: Aspect ratio equal to 2.5:1, porosity equal to 40 % (substrate medium size equal to 3.5 mm), depth equal to 0.27 m and HLR equal to 20-27 mm/day. Temperatures around 5-15°C (mild weather) are also recommended.

1 Introduction and goals

From 2001 to 2003, the Environmental Engineering Division of the Hydraulics, Coastal and Environmental Engineering Department of UPC Barcelona Tech carried out data recording on their horizontal subsurface flow constructed wetlands (from now on, CW SSF) in the municipality of Les Franqueses, Spain (García et al., 2005). The CW imitates the filtering and pathogen deactivation mechanisms of natural wetlands. This team's goal was to find a suitable CW SSF for the treatment of municipal wastewater.

There were eight types of CW SSFs under study. All of them sized 54-56 m^2. Common reed (*Phragmites australis*) was grown in them. The CW SSFs were connected in parallel to a wastewater source (fig. 1) and each one received ideally the same discharge along the observation time. Such waste water was Les Franqueses municipal disposed water after being treated by an Imhoff tank. The CW SSFs were named after the letters: A, B, C and D (see table 1), depending on their aspect ratio (ratio "length : width"). The CW SSFs named A had an aspect ratio of 1:1, the ones named B, of 1.5:1, the ones named C, of 2:1, and the ones named D of 2.5:1 (see fig. 1). On the other hand, CW SSFs A's water depth were 0.5 m, CW SSFs B's were 0.5 m in 2001 and 2002 and changed to 0.27 m in 2003, CW SSFs C's were 0.5 m and CW SSF D's were 0.27 m. Taking the same CW SSFs, they are also split into groups 1 and 2. Different substrate medium sizes (D_{60}) and porosities (ϕ) were associated bi-univocally with each of these two groups: group 1 had a substrate (coarse granitic gravel) medium size of 10 mm and a porosity of 39%, while group 2 had a substrate (small granitic gravel) medium size of 3.5 mm and a porosity of 40%. Independently, four different hydraulic load rates (from now on, HLR) were applied in order to test HLR effects on CW SSF water depuration: 20, 27, 36 and 45 mm/day. Therefore, there seems to be five factors but only four are effectively considered herein. Reasons are given later. Affluence water temperature and effluence water temperature are also taken into account as covariates. Moreover, data is tested for other factors such as season, rain and age in a further study. Some conclusions in García et al. (2005) are that the most suitable water depth

and substrate type are 0.27 m and fine gravel (D_{60}=3.5 mm, ϕ=40%). More conclusions are that HLR determines results and aspect ratio does not. These statements are examined after applying Compositional Data Analysis to the data used in this previous experiment.

Section 2 resumes the data analysis in García et al. (2005). Section 3 is a summary of the statistical methods used in the re-examination carried out here. Section 4 presents and discusses the results, and Section 5 displays the conclusions.

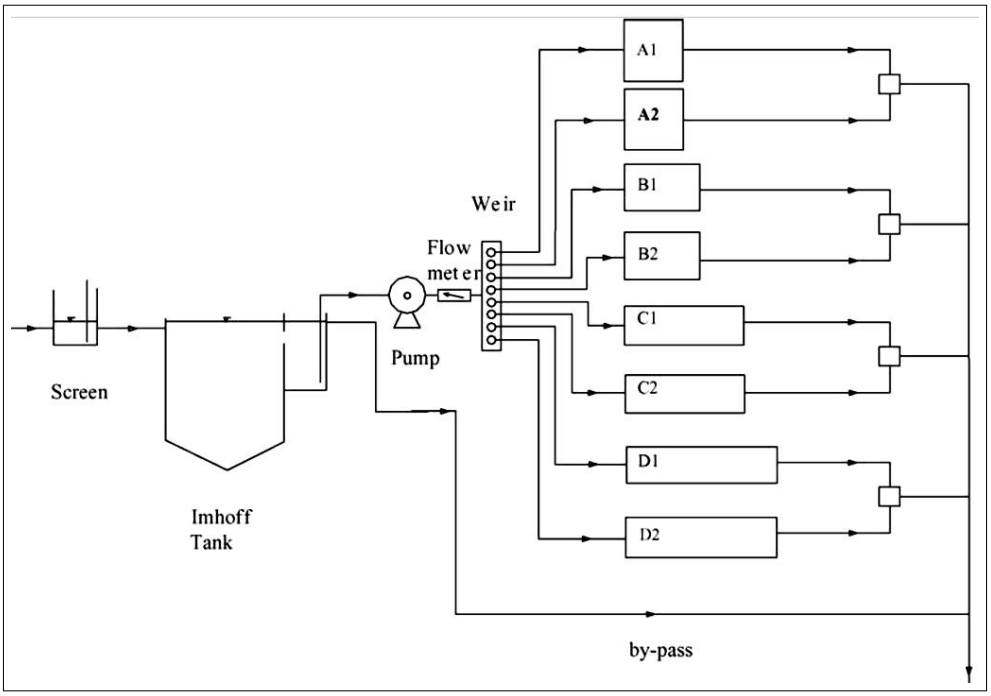


Figure 1: Outline of the experimental CW SSFs. Courtesy of Dr. Joan García

Factor	Number of levels	Levels
Aspect ratio	4	A (1:1), B(1.5:1), C(2:1), D(2.5:1)
Substrate medium size / porosity	2	3.5 mm (40 %), 10 mm (39 %)
Water depth	2	0.27, 0.5 m
HLR	4	20, 27, 36, 45 mm/day

Table 1: Summary of different levels of the factors

2 Data

The raw data set kindly provided by Dr. J. García has the following information:

- factors: sampling, date, year, bed type (A1, A2, B1, B2, C1, C2, D1 and D2), flow (programmed discharge, in m^3/day), real flow (real discharge measured with a flowmeter, in m^3/day), HLR (value obtained dividing flow by the surface area of the CW SSFs, in mm/day), bed aspect ratio, substrate medium size (medium size of the substrate material used, in mm), porosity (in %), water depth (in meters), season, rain condition (whether it rained).
- variables, from both the affluence and the effluence: COD, BOD_5, NH_3, PO_4^{3-} (all expressed in mg/l), temperature (in Celsius degrees) and pH (APHA-AWWA-WEF, 2005).

As grains of D_{60}=3.5 mm always create an environment of 40% of porosity and grains of D_{60}=10 mm always create an environment of ϕ=39%, the substrate medium size and the porosity are considered

as one single joint factor when analyzing the data. On the other hand, HLR=27 mm/day was only applied 16 times while HLR=20 mm/day was applied 115 times: there is a significant number of samples corresponding to a HLR of 20 mm/day compared to those of 27 mm/day. Since they are relatively close numbers, compared to 36 or 45 mm/day, it has been preferred to join 20 and 27 mm/day as a single factor level. Each composition is represented as follows: COD is COD, BOD_5 is BOD, NH_3 is NH3, PO_4^{3-} is PO4, H^+ is H and water plus other components (obtained as 10^6 minus COD, BOD_5, NH_3, PO_4^{3-} and H^+) is O. Affluence variables will bear the prefix "A-" and effluence variables will bear the prefix "E-". If perturbation and closure (Pawlowsky-Glahn et al., 2011) is applied to the compositions, i.e. ACOD⊖ECOD, new compositions of this nature can better tell the transformation suffered by compositions through the CW SSFs. They can also be written with the prefix "perturb-". For example: ACOD⊖ECOD can be found written as perturbACOD in some figures.

It is important to interpret missing data correctly. For instance, in the present case, they are not under-detection-limit values. Here, they are simply data points which were not measured during the samplings. If these data points were considered as under detection-limit, a significant error related to outliers would be introduced. In figure 2, there are two extra visible clusters that do not exist in the real experiment: one on the upper right hand corner and another one on the left of the center of the axes. Robust methods would be needed (Filzmoser and Templ, 2011) and their application would assure more reliable results. These missing data cases are removed in the following analysis, and apparently, standard methods work properly in most cases.

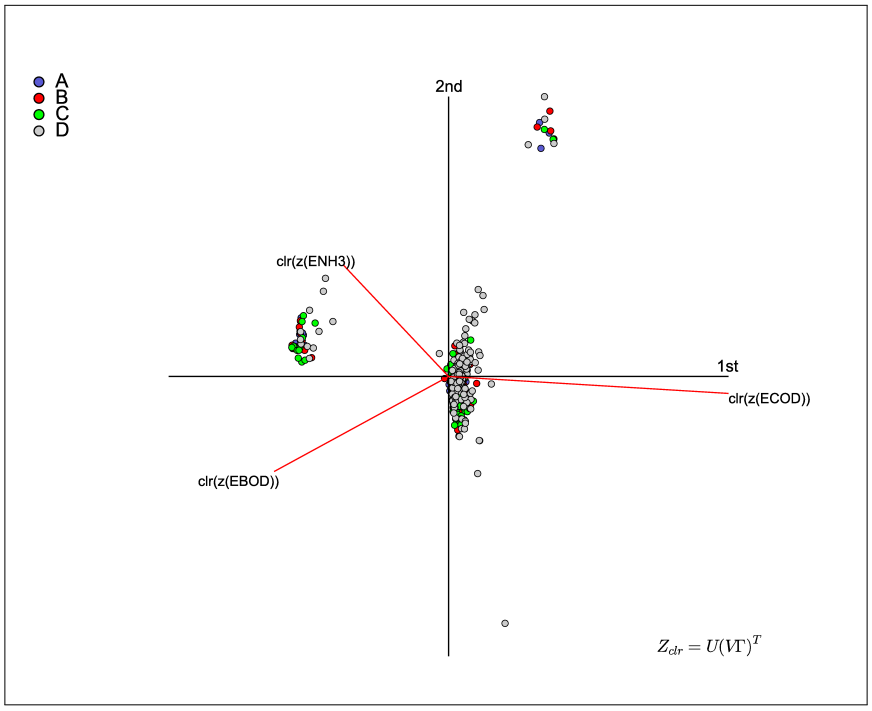


Figure 2: Biplot of ECOD, EBOD and ENH3. Data is grouped by aspect ratios: A (1:1), B (1.5:1), C (2:1) and D (2.5:1). Missing data, interpreted as under-detection-limit observations, are imputed. Two extra visible clusters that do not exist in the real experiment appear: one on the upper right hand corner and another one on the left of the center of the axes.

The variables chosen are: ACOD⊖ECOD, ABOD⊖EBOD, ANH3⊖ENH3, APO4⊖EPO4, AH⊖EH and AO⊖ EO. The geometric centers of the first four are listed on Table 3. There are a total of 824 samples. After ruling out the rows with missing data points, there are 410 samplings left.

A sequential binary partition (Egozcue and Pawlowsky-Glahn, 2005) is selected to apply the ilr transformation (Egozcue et al., 2003) to the the chosen variables (see Table 4). First, as PO_4^{3-} was not taken into account in the beginning, it is set against all other components. The reason why it was

	clr(perCOD)	clr(perBOD)	clr(perNH3)	clr(perPO4)	clr(perH)	clr(perO)	Cum.Prop. Exp.
PC1	0.1588	0.6258	0.2915	-0.1339	-0.6057	-0.3366	0.5366
PC2	0.0007	0.5660	-0.5351	-0.4428	0.4430	-0.0318	0.7741
PC3	0.5709	-0.0471	-0.6424	0.3908	-0.3224	0.0502	0.8729
PC4	-0.6913	0.3442	-0.2201	0.5681	-0.1276	0.1267	0.9582
PC5	0.0648	-0.0247	0.0304	0.3793	0.3873	-0.8369	1.0000

Table 2: Principal components (PC) of biplot after removing missings. An example to explain what the columns are: "perCOD" is "perturbCOD".

	$\mu(ACOD \ominus ECOD)$	$\mu(ABOD \ominus EBOD)$	$\mu(ANH3 \ominus ENH3)$	$\mu(APO4 \ominus EPO4)$
Aspect ratio				
A (1:1)	0.2789	0.2738	0.1402	0.1085
B (1.5:1)	0.2729	0.2858	0.1394	0.1060
C (2:1)	0.2730	0.2814	0.1498	0.1063
D (2.5:1)	0.2607	0.4207	0.1483	0.0644
Porosity (in %)				
39	0.2677	0.3006	0.1405	0.1023
40	0.2809	0.3268	0.1515	0.0889
Depth (in m)				
0.27	0.2606	0.3938	0.1467	0.0754
0.5	0.2783	0.2727	0.1436	0.1070
HLR (in mm/day)				
20	0.2736	0.3414	0.1432	0.0887
27	0.3222	0.3054	0.1345	0.0761
36	0.2632	0.3106	0.1483	0.0991
45	0.2928	0.2667	0.1485	0.1062

Table 3: Means of data classified by factors.

ISBN: 978-3-200-03103-6

not considered is that it does not provide as much additional information as the other components do, since it is related to BOD_5 and its clr variance is lower than the latter. Second, water and other elements is set against COD, BOD_5, NH_3 and H^+. Third, oxidizable matter, represented by COD and BOD_5, is set against nitrogen (NH_3) and hydrogen. Finally, for each given pair, each member is set against one another. Returning to the chosen variables, a centered log-ratio transformation (clr) is applied. The clr variables are used to build compositional biplots (Aitchison and Greenacre, 2002) which provide interesting information. For instance, according to the biplots, the variables are well chosen, as their rays are spread all over the real space, rather than concentrating in one single area of it (see figs. 4, 5 and table 2).

Balance	ACOD⊖ ECOD	ABOD⊖EBOD	ANH3⊖ ENH3	APO4⊖EPO4	AH⊖ EH	AO⊖ EO
1	+	+	+	−	+	+
2	+	+	+		+	−
3	+	+	−		−	
4	+	−				
5			+		−	

Table 4: Sequential binary partition of ACOD⊖ ECOD, ABOD⊖EBOD, ANH3⊖ENH3, APO4⊖EPO4, AH⊖EH and AO⊖EO. This table will also be applied to affluence and effluence data.

3 Methods

Clr and ilr transformations are applied to the data used in García et al. (2005). Compositional statistics summaries are carried out on the compositional variables and classical statistics summaries are carried out on the ilr transformed variables. Also, dendrograms and biplots are plotted. A linear model is used in order to test the equality of means of data grouped by different levels of a factor. Along with the dendrograms and biplots, conclusions can be drawn for the question of which features are the most suitable for a CW SSF. The following two subsections are merely a reminder of the basic statistics applied here.

3.1 ANOVA (ANalysis Of VAriance)

ANOVA is the test of equality of the means of g groups (being $g \geq 2$) on the single response variable (RV). Being the hypotheses

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_g$$

versus H_1 : at least two μ's are unequal.

When the effects of one single factor is examined, the test is called *one-way*. If more factors are considered, the analysis is *multiple-way*. The case studied corresponds to a multiple-way analysis. For simplicity, an example of the model of a two-way ANOVA with interactions follows. Consider a factor F_1 including levels A, B and C and a factor F_2 including levels M and N, then the linear model to be fitted is

$$y_i = \mu_{AM} + \mu_B I_B + \mu_C I_C + \mu_N I_N + \mu_{BM} I_B I_M + \mu_{CM} I_C I_M + \mu_{AN} I_A I_N + \mu_{BN} I_B I_N + \mu_{CN} I_C I_N + \epsilon_i$$

where μ_j, μ_k, μ_{jk} are means, I_j, I_k are indicators and ϵ_i is the total error, for $j = A, B, C$ and $k = M, N$. The testing approach to solve the question of equality of means is based on the decomposition of the sum of squares

$$SS_T = SS_B + SS_W,$$

where SS_T is the total sum of squares, SS_B is the sum of squares between the means of the groups and SS_W is the sum of squares within the groups. The statistic for testing the null hypothesis is the following F-ratio

$$F = \frac{\text{mean}\,(SS_B)}{\text{mean}\,(SS_W)} = \frac{[SS_B/\,(g-1)]}{[SS_W/\,(n-g)]} \sim F_{g-1,n-g}.$$

The standard assumptions of ANOVA are: the residuals are independent, normally distributed with equal variance (homoskedacity). If the assumptions are violated, robust models can be used in order to obtain trustworthy results (Filzmoser and Templ, 2011).

3.2 MANOVA

Compositional Multivariate ANalysis Of VAriance (MANOVA) can be inferred from ANOVA and classical statistics. It examines the mean differences across several treatments or factors when more than one response variable are considered simultaneously. That is, MANOVA is essentially an analogy of ANOVA with $p \geq 1$ response variables. Therefore, all the concepts related to it and the assumptions when applying it are the $p \geq 1$ versions of the ones for ANOVA.

It is assumed that k independent random samples of size n are obtained from p-variate normal populations with equal covariance matrices. Mean vectors are to be compared for significant differences. The hypothesis is, analogous to ANOVA, $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ versus H_1 : at least two μ's are unequal.

MANOVA can also be one-way or multiple way. The most usual MANOVA statistics are Wilk's Lambda, Hotelling's Trace, Pillai's Trace and Roy's Largest Root. For each case there is a different criteria to tell if the null hypothesis is rejected. Remarkably, MANOVA is, in its turn, a particular case of linear model. Therefore, when there is interest in analyzing the interaction of the given factors (CW SSF aspect ratio, substrate medium size / porosity, water depth and HLR), and also using covariates (temperature), MANOVA must be ruled out in favor of a linear models, as temperature presents many more levels that would yield an *error* response in conventional statistic softwares if MANOVA was used.

4 Results and discussions

Affluence and effluence data are set apart from each other, being both affluence and effluence balances (Egozcue et al., 2003) obtained from the partition in table 4. In this manner, it is possible to study the composition of affluents independently from the composition of effluents. It is observed that along the CW SSFs the proportions of pollutants decrease for all levels of all factors. Therefore, what we had was an acceptable wastewater treatment that did actually reduce the pollutants, which seems trivial but may not be achieved in all experimental water treatment methods. Meanwhile, balance standard deviations increase. The increase in variance is a sign of good functionality. It also seems that dependencies grew between the components along the CW SSFs as covariances of all ilr variables increase. corr(ilr1, ilr3), corr(ilr1,ilr4), corr(ilr3, ilr4) and corr(ilr4, ilr5) increased inside the CW SSFs, as well. Higher correlation means that the amounts of one component can be inferred from the amounts of another one. The latter two correlations jointly mean the one between organic-matter concentration and nitrogen concentration. The reason why covariances and correlations increased between phosphates and organic matter is that the organic matter, which includes microorganisms, can take phosphate as a nutrient. The reason for which covariances and correlations increase between the organic matter and the nitrogens is that the anoxic environment, which is good for nitrification, is inappropriate for organic matter removal, leading to a certain dependency between organic matter and nitrogens.

Focusing on the transfer compositional data (i.e. ACOD$\ominus$ ECOD), CW SSFs with an aspect ratio of 2.5:1 reduce more pollutants (see fig. 3 and tables 5 and 6), particularly more biochemically oxidizable matter and nitrogens, as $\mu(ilr2)$ is relatively higher compared to other aspect ratios, indicating higher pollutant removal rate; $\mu(ilr3)$ is relatively higher, indicating more oxidizable matter removal; $\mu(ilr4)$ is relatively lower, indicating higher ABOD$\ominus$ EBOD removal, and $\mu(ilr5)$ is relatively higher, indicating higher ANH3$\ominus$ENH3 removal. Compositional biplots of data from different aspect ratio CW SSFs support these results. All graphs will be available in a future minor thesis from the author. The reason why longer CW SSFs are more efficient in organic matter and nitrogens removal might be that elongated shapes prevent side corner eddies that do not allow biochemical reactions to be completed.

Higher porosity helps reducing pollutants, particularly organic matter and nitrogens, as $\mu(ilr2)$ is higher for ϕ=40% than for ϕ=39%, indicating greater pollutant removal rate; $\mu(ilr3)$ is higher, indicating more oxidizable matter removal and $\mu(ilr5)$ is higher, indicating more ANH3⊖ENH3 removal. García et al. (2010) commented that CW SSFs do not depend on adsorption for waste removal from water. Hence, along with results stated above, voids are preferred over surfaces. Furthermore, the substrate material used in García et al. (2005) is granitic gravel, which is a dense material. A porous material would provide additional porosity.

Figures 4 and 5 show that shallower CW SSFs are more successful in removing pollutants, particularly organic matter and nitrogens. Shallow CW SSFs create a more suitable re-dox environment for organic-matter removal, the aerobic one (García et al., 2010). On the other hand, such a relatively aerobic environment helps completing the nitrification-denitrification cycle needed to completely remove nitrogens from the system, as CW SSFs usually hold anaerobic environments that only allow for nitrification, unlike vertical subsurface flow constructed wetlands (García-Serrano and Corzo-Hernández, 2008).

Graphs plotted from data grouped by HLR do not give clear information, as water had been driven certain distances inside pipes before reaching the CW SSFs, but some conclusions can be drawn. For instance, 20-27 mm/day is best at removing pollutants, specially chemically oxidizable matter and nitrogens.

On the other hand, phosphates seem to behave exactly the opposite way than other pollutants. Their removal need shorter CW SSFs, which ϕ should be 39%, water depth should be 0.5 m and HLR should be 45 mm/day. Phosphate can come mainly from inorganic sources, unlike organic matter or nitrogens. This fact can be the reason for this disparity in result.

Finally, further analyses show that when it does not rain, CW SSFs seem to achieve more pollutant (biochemically oxidizable matter and nitrogens) removal and that as the CW SSFs age, they seem to operate better. Rain conditions determine components concentration and the interaction between overall pollutant concentration and organic matter are well observed. Aging, on the other hand, allows pollutant processing microorganisms and macroorganisms to be grown in the CW SSFs. It can be observed that the conclusions about the CW SSF porosity (or substrate medium size), depth and age are the same as in García et al. (2005). However, the present study states that aspect ratio is very important for the CW SSFs while García stated that it was not.

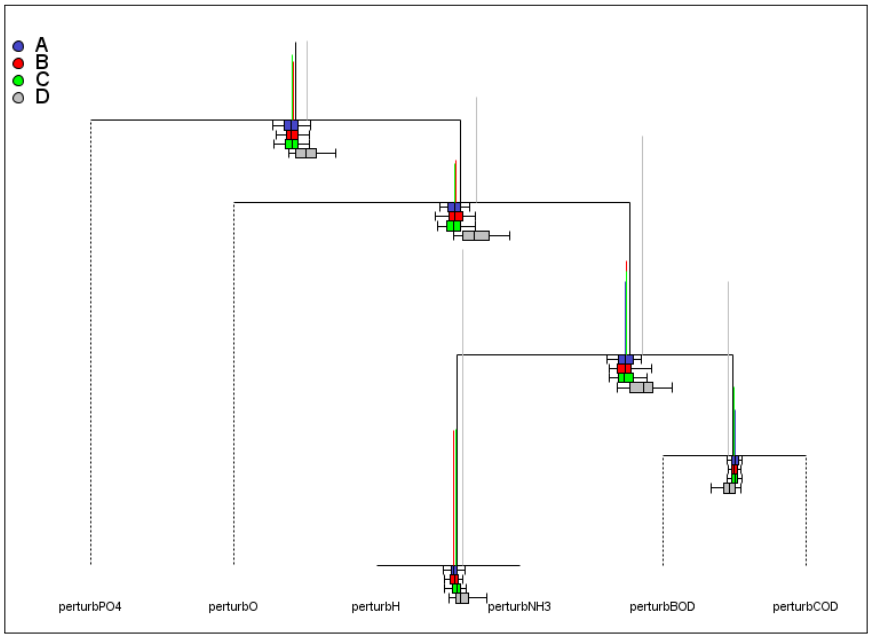


Figure 3: Dendrogram of data grouped by aspect ratio: A(1:1), B(1.5:1), C(2:1) and D(2.5:1).

Being ilr1, ilr2, ilr3, ilr4 and ilr5 the balances obtained from the partition in table 4, MANOVA

Balance 1	Balance 2	Balance 3	Balance 4	Balance 5
0.4432	0.5796	1.0183	-0.0946	0.4532

Table 5: Mean of dendrogram

Balance 1	Balance 2	Balance 3	Balance 4	Balance 5
0.1013	0.1057	0.1964	0.1298	0.2719

Table 6: Variance of dendrogram

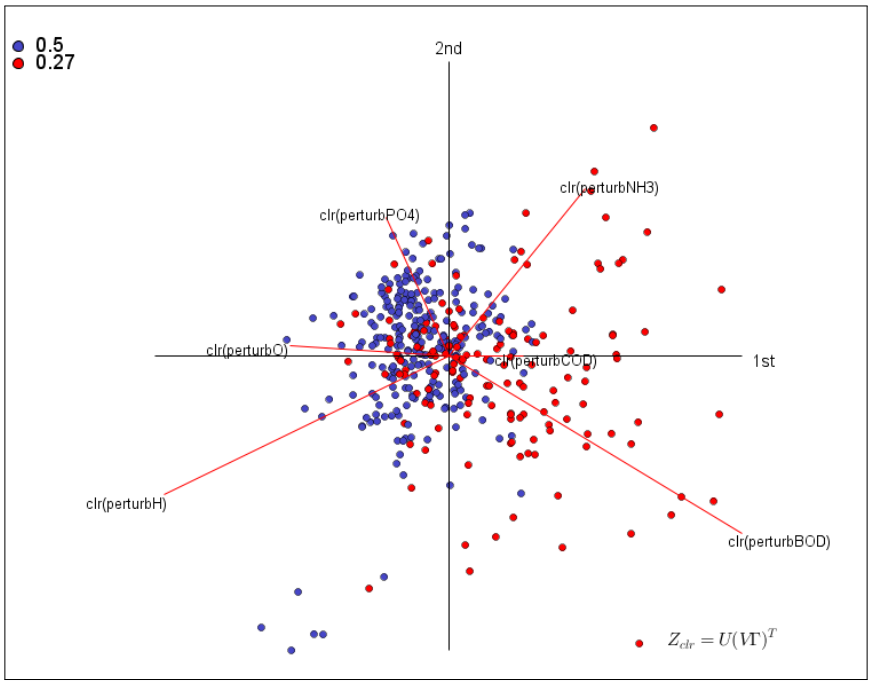


Figure 4: Biplot of data grouped by water depth: 0.5 and 0.27 m. PC1 against PC2.

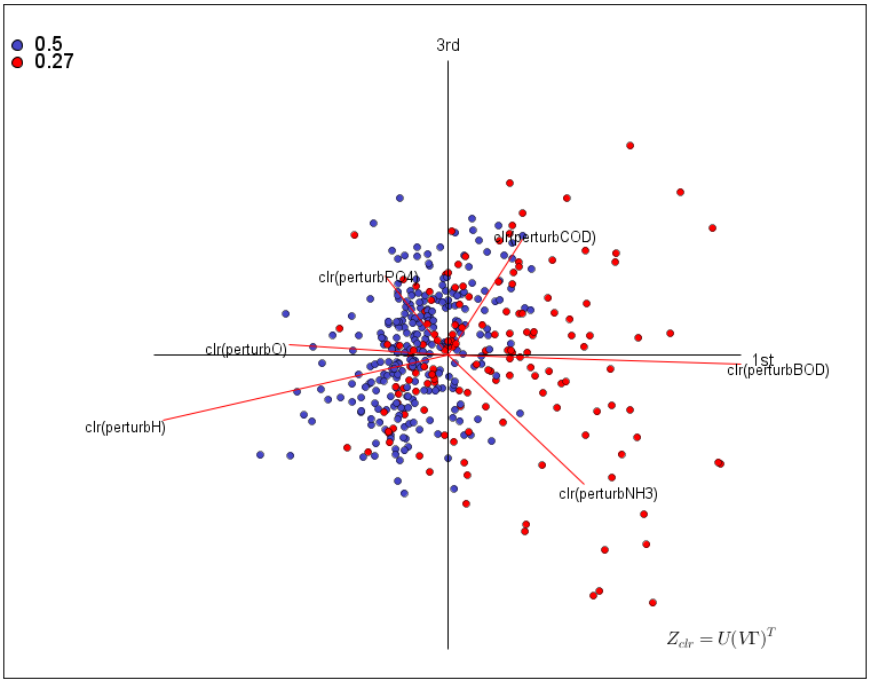


Figure 5: Biplot of data grouped by water depth: 0.5 and 0.27 m. PC1 against PC3.

techniques are applied to them (García-León and Lin, 2011), the Wilks statistic is used. The ilr transformed variables are previously tested for MANOVA assumptions. The variables fulfill the hypotheses well except in some cases as, for instance, homoskedacity of ilr5 for porosity. This assumption violation might be due to the lack of porosity variety. The two kinds of substrate, coarse gravel (D_{60}=10 mm and ϕ=39%) and fine gravel (D_{60}=3.5 mm and ϕ=40%) might be excessively different and a wider range of levels with more gradually changing variances might be needed for a more fruitful study. Robust models are used in this and other cases where assumptions are violated.

Given a significance level of 0.05, null hypotheses are rejected for p-values below significance level. Ilr2 is tested to be different for aspect ratio, D_{60}/porosity, depth, HLR and temperature, yielding low p-values (see table 7). Therefore, all these factors are decisive for overall water treatment efficiency and the most suitable features are: aspect ratio equal to 2.5:1, ϕ=40% (D_{60}=3.5 mm), depth=0.27 m, HLR=27 mm/day and temperature around 6°C.

The last three ilrs are tested significantly different for different aspect ratios. The longer the CW SSF, the best organic matter and nitrogens removal and worst phosphates removal (see figure 3 and table 3). p-value of testing ilr1 for H0: " results for an aspect ratio of 2:1 and an aspect ratio of 2.5:1 are equal" is 3.245×10^{-13}. p-value of testing ilr3 for H0: "results for an aspect ratio of 2:1 and an aspect ratio of 2.5:1 are equal" is 5.726×10^{-12}, while null hypothesis is not rejected for 1:1 or 1.5:1 Vs. 2:1. Similarly, p-value for ilr5 is below significance level for and aspect ratio of 1:1 or 1.5:1 Vs. 2:1 and for 2:1 Vs. 2.5:1. Hence, the most suitable aspect ratio is 2.5:1.

The substrate medium characteristics are also decisive in nitrogens and organic matter removal but not in reducing phosphates. For H0: "results for ϕ=39% are equal to results for ϕ=40%", p-value for ilr1 is 0.109, for ilr3 is 0.000105 and for ilr5 is 0.001387. A substrate medium size of 3.5 mm and 40% porosity is the most suitable choice.

Ilr 1 is tested to be different for 20-27 mm/day versus 36-45 mm/day. Consequently, phosphates are best removed for HLR=36-45 mm/day. Ilr3 is tested to be different for 20-27 mm/day versus 36-45 mm/day. This means that for one of these HLR intervals either the organic matter elimination rate is higher or the nitrogen removal rate is lower. According to table 3, organic matter is best removed for HLR=20 to 27 mm/day. Nitrogens removal does not depend on HLR.

Performance depends on depth. For H0: "results for a water depth of 0.27 m are equal to results for a water depth of 0.50 m", p-value for ilr1 is 6.423×10^{-11}, p-value for ilr3 is 2.519×10^{-13} and

	p-values				
H0	Balance 1	Balance 2	Balance 3	Balance 4	Balance 5
Aspect ratio					
A Vs. B	0.4457	0.2978	0.2766	0.2417	0.8597
B Vs. C	0.5432	0.2154	0.2984	0.5591	0.01007
C Vs. D	3.245×10^{-13}	2.2×10^{-16}	5.726×10^{-12}	3.496×10^{-6}	0.0001971
ϕ (%)/$D_{60}(mm)$					
39 Vs. 40	0.109	0.02991	0.000105	0.9477	0.00387
Depth (m)					
0.27 Vs. 0.50	6.423×10^{-11}	$<2.2\times10^{-16}$	2.519×10^{-13}	4.564×10^{-13}	2.44×10^{-5}
HLR (mm/day)					
20 or 27 Vs. 36	0.01798	0.006879	0.002294	0.01207	0.2116
36 Vs. 45	0.1982	0.001951	0.9023	4.117×10^{-5}	0.5717
Temperature					
ilr$\sim ATemp + ETemp$	$<2.2\times10^{-16}$	0.002771	2.492×10^{-12}	$<2.2\times10^{-16}$	4.528×10^{-16}

Table 7: p-values of one-way ANOVAs and some multiple-way ANOVAs carried out in this analysis. These p-values are meant to outline the whole analysis results. ATtemp is the affluence temperature and ETemp is the effluence one.

p-value for ilr5 is 2.44×10^{-5}. Therefore, water depth around 0.27 m is the best choice to minimize organic matter and nitrogens presence in the effluence but higher concentrations of phosphates are expected.

Temperature plays a key role in water treatment. A simple linear regression show that the optimum temperature for organic matter removal is around 25.5°C, as p-value of ilr3 for the linear regression ilr$\sim ATemp + ETemp$ is 2.492×10^{-12} and ilr3 increases with temperature rise. However, phosphate and nitrogens removal decrease with temperature increase, as p-value of ilr1 for the linear regression ilr$\sim ATemp + ETemp$ is under 2.2×10^{-16} and p-value of ilr5 is 4.528×10^{-16} and the same linear regressions show that these values decrease with temperature rise.

Finally, one-way MANOVA tests of ilr1 through ilr5 yield p-values below significance level for the factors aspect ratio, ϕ/D_{60}, depth and HLR. Also, as ilrs are tested for the interactions of aspect ratio, porosity, depth and temperature, it can be stated that there is synergy between, for instance, overall pollutant concentration and organic matter removal. To a certain limit which is not quantified here, high concentrations of nutrients help CW SSFs microorganisms digest organic matter. This explains why the CW SSFs seem to perform better in non-rainy days.

Finally, the following log-contrast is used as a criterion for the CW SSF efficiency

$$CWE = ln\frac{(ACOD \ominus ECOD)^3(ABOD \ominus EBOD)^3(ANH3 \ominus ENH3)^2}{(AO \ominus EO)^8} .$$

The Constructed Wetland Efficiency (CWE) index describes in a simple manner the functionality of the CW. The main goal has been set as removing COD and BOD_5 (organic matter), both raised to the 3rd power, and the secondary goal has been set as nitrogens removal (NH_3), raised to the 2nd power. The product of these three is divided by the change in proportion of water to pollutants, raised to the 8th power. Regard that the power of the numerator, 3+3+2=8 is equal to the power of the denominator, in order to cancel the units. Logaritm is applied to this index as it can only take positive nonzero values. For the linear regression CWE$\sim Aspect.Ratio * Porosity * Depth * HLR * (ATemp + ETemp)$, CWE does depend on aspect ratio (p-value$<2.2\times10^{-16}$), porosity (p-value=4.265×10^{-12}), depth (p-value=0.000379) and HLR (p-value=1.123×10^{-10}). It is also influence by the interaction of aspect ratio and porosity (p-value=2.912×10^{-10}), aspect ratio and HLR (p-value=0.007224), aspect ratio and temperature, depth and temperature and HLR and temperature. Here, it is obvious that aspect ratio and temperature could not be ignored as factors. CWE is maximum for elongated CW SSFs (aspect ratio of 2.5:1), which ϕ=40 % (D_{60}=3.5 mm), water depth is 0.27 m, HLR is 20-27 mm/day and temperature is around 6°C. According to CWE, the maximum efficiency had been in November 8th 2003, when aspect ratio was 2.5:1, ϕ=40% (D_{60}=3.5 mm) , depth=0.27 m, HLR=20 mm/day

and temperature was around 16°C. In this case, the result is consistent with previous conclusions. About temperature, although organic matter removal increases with its rise, the overall result from CWE and from the whole compositional analysis carried out is that CW SSFs are most efficient for temperatures around 5-15°C.

5 Conclusions

Significative ilr-mean differences have been found between CW SFF configurations using ANOVA and MANOVA analysis. Also the effect of temperature has been found significative in a regression analysis. The influence of aspect ratio is found to be significative although it was not found in the previous analysis (García et al., 2005) using non-compositional analysis. A log-contrast applied additionally verifies these results. Most efficient CW SSFs for overall, organic matter and nitrogens removal have the following features:

- Aspect ratio equal to 2.5:1
- Porosity equal to 40 % and substrate medium size equal to 3.5 mm
- Depth equal to 0.27 m
- HLR equal to 20-27 mm/day

Also, temperatures around 5-15°C (mild weather) is also recommended.

6 Acknowledgements

I would like to thank Prof. J. J. Egozcue for his patient tutoring and his funding which allows me to attend CoDAWork'13. I would also thank Dr. J. García for kindly providing the data from his experiment.

References

Aitchison, J. and J. Greenacre (2002). Biplots of compositional data. *Applied Statistics* (51), 375–392.

APHA-AWWA-WEF (2005). *Standard methods for the examination of water & wastewater* (21st ed.). American Public Health Association.

Egozcue, J. J. and V. Pawlowsky-Glahn (2005, October). Groups of parts and their balances in compositional data analysis. *Mathematical Geology 37*(7), 825–827.

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

Filzmoser, P., H. K. and M. Templ (2011). Robust compositional data analysis. In *Proceedings of the 4th International Workshop on Compositional Data Analysis (2011).*

García, J., P. Aguirre, J. Barragán, R. Mujeriego, V. Matamoros, and J. M. Bayona (2005). Effect of key design parameters on the efficiency of horizontal subsurface flow constructed wetlands. *Ecological Engineering*, 405–418.

García, J., D. P. L. Rousseau, J. Morató, E. Lesage, V. Matamoros, and J. M. Bayona (2010). Contaminant removal processes in subsurface-flow constructed wetlands: A review. *Critical Reviews in Environmental Science and Technology 40*(7), 561–661.

García-León, M. and J. Lin (2011). Testing water pollution in a two layer aquifer. In *Proceedings of the 4th International Workshop on Compositional Data Analysis (2011).*

García-Serrano, J. and A. Corzo-Hernández (2008). Depuración con humedales construidos: guía práctica de diseño, construcció y explotación de sistemas de humedales de flujo subsuperficial. Technical report, DEHMA.

Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2011). Lecture notes on compositional data analysis.

Have you got things in proportion? A practical strategy for exploring association in high-dimensional compositions

D. LOVELL[1], V. PAWLOWSKY-GLAHN[2] and J. J. EGOZCUE[3]

[1] CSIRO Mathematics, Informatics, and Statistics, Canberra, Australia, David.Lovell@csiro.au
[2] Dept. Informatics and Applied Mathematics, U. de Girona, Spain
[3]Dept. Applied Mathematics III, U. Politécnica de Catalunya, Barcelona, Spain

Abstract

This paper is motivated by the need to apply compositional data analysis to molecular bioscience. Awareness is gradually growing that most bioscience measurement processes yield only relative information, and that these data require appropriate analysis and interpretation. We are keen to ensure that practical compositional methods exist to provide well-principled alternatives to the potentially misleading correlation-based methods that are generally used in bioscience today.

In common with the companion paper in these proceedings, the strategy we present is based on finding pairs of components that are proportional, or very nearly so. We do this here using the *standardised major axis* (SMA) (Warton et al., 2006) to provide estimates of the slopes of the bivariate relationships between components. This leads to a factorization of logratio variance $\mathrm{var}(\log(x/y))$ in which the "goodness of fit to proportionality between x and y" is clear. We show how this goodness of fit can then be used as the basis for multivariate analysis strategies familiar in the molecular biosciences, including network inference, heatmaps and hierarchical clustering. Also we show why we prefer goodness of fit over hypothesis testing in this application.

We explore the usefulness, strengths and limitations of this approach by using it to find proportional sets of components in yeast gene expression data where the levels of 3031 messenger RNAs were observed in a 16-point time course.

1 Introduction

In molecular bioscience, measurements of relative abundance are, well... *abundant*. Yet appreciation of the need to analyse and interpret these data differently to measurements of absolute abundance is *scarce*. In keeping with experiences in the geosciences (Buccianti et al., 2006), we believe it will take patience and persistence for concerns about the analysis of compositional data to influence general practice in bioscience (Lovell et al., 2011).

Still, awareness *is* growing: Lovén et al. (2012) recently highlighted the potential to misinterpret relative abundance in gene expression measurements; Faust et al. (2012) and Friedman and Alm (2012) emphasise compositional issues in metagenomics data and propose algorithms towards ameliorating these issues, acknowledging—but not embracing—Aitchison's (1986) logratio approach.

For the analysis of relative abundance data to be put on a sound footing in molecular bioscience, compositional methods must be proposed that can be readily incorporated into the canon of multivariate methods familiar to molecular bioscientists, including network inference (Bansal et al., 2007), heatmaps and hierarchical clustering (Eisen et al., 1998). This is our aim in this paper; we focus our efforts on logratio variance and use *proportionality* as a measure of association for relative abundance data, rather than the current, oft misguided workhorse of molecular bioscience—correlation (Lovell et al., 2013).

2 Where's proportionality in logratio variance?

The logratio variance of two components x and y of a composition[1] is $\mathrm{var}(\log(x/y))$ and "...should form a useful tool for investigating the pattern of variability of a composition" (Aitchison, 1986, Section 4.3). Clearly, when x and y are proportional to one another $\mathrm{var}(\log(x/y)) = 0$. However, when x and y are not exactly proportional, "it is hard to interpret as it lacks a scale. That is, it is

[1]To simplify notation we will use single letters to denote components where practicable.

unclear what constitutes a large or small value... (does a value of 0.1 indicate strong dependence, weak dependence, or no dependence?)" (Friedman and Alm, 2012).

We have found a way to factorize logratio variance so that it is more interpretable. Ironically, this approach relies upon correlation, and an approach to bivariate line fitting that has some very useful properties—the standardised major axis.

2.1 One line to rule them all: the standardised major axis

"Fitting a line to a bivariate dataset can be a deceptively complex problem" write Warton et al. (2006) who suggest that, when interest focuses on the slope of the relationship between two variables, *major axis* (MA) or *standardised major axis* (SMA) lines should be used rather than ordinary least squares regression (which will underestimate the slope of the line of best fit). We concentrate on the SMA approach because its slope estimate has a clear relationship to logratio variance.

The SMA estimate of slope of random variables b on a is

$$\hat{\beta}(a,b) = \text{sign}(s_{ab})\frac{s_b}{s_a}$$

where s_{ab} is the sample covariance of a and b, and s_a^2 and s_b^2 are the sample estimates of the variances of a and b, respectively. Note that this slope is symmetric: the slope of $a \sim b$ is the inverse of the slope of $b \sim a$.

If we set $a = \log x$ and $b = \log y$ then

$$\hat{\beta}^2(\log x, \log y) = \frac{\text{var}(\log y)}{\text{var}(\log x)} \tag{1}$$

2.2 Correlation rides again, with determination

The squared sample correlation coefficient of random variables a and b is known as the *coefficient of determination* and estimates the strength of the linear relationship between them:

$$r^2(a,b) = \frac{s_{ab}^2}{s_a^2 s_b^2}$$

If we set set $a = \log x$ and $b = \log y$ then

$$r^2(\log x, \log y) = \frac{\text{cov}(\log x, \log y)^2}{\text{var}(\log x)\text{var}(\log y)} \tag{2}$$

2.3 A variation on logratio variance

To make use of Equations 1 and 2 we expand logratio variance:

$$\begin{aligned} \text{var}(\log(x/y)) &= \text{var}(\log x - \log y) \\ &= \text{var}(\log x) + \text{var}(\log y) - 2\text{cov}(\log x, \log y). \end{aligned} \tag{3}$$

Now we are in a position to combine Equations 1 to 3. For ease of notation we define

$$A \triangleq \text{var}(\log x), \qquad B \triangleq \text{var}(\log y), \qquad C \triangleq \text{cov}(\log x, \log y),$$

which means that

$$\begin{aligned} r^2(\log x, \log y) &= \frac{C^2}{AB} \\ \hat{\beta}^2(\log x, \log y) &= \frac{B}{A} \\ \text{var}(\log(x/y)) &= A + B - 2C. \end{aligned}$$

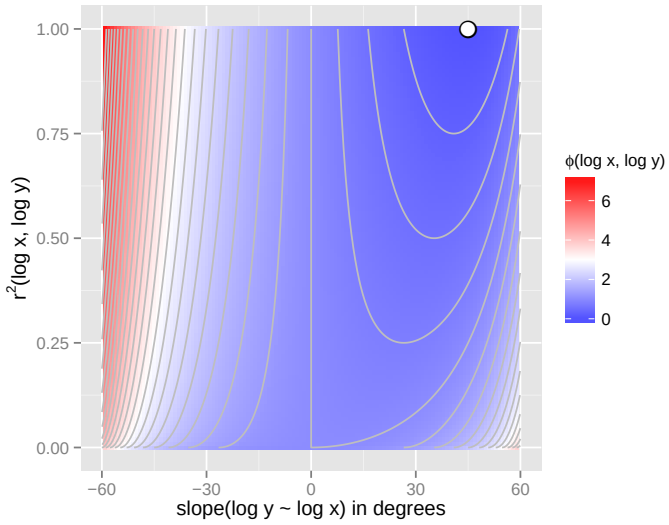


Figure 1: $\phi(\log x, \log y)$ as a function of the slope and coefficient of determination of the standard major axis of $\log y$ versus $\log x$. The grey lines show the contours of $\phi(\log x, \log y)$ in increments of 0.25. The hollow dot shows the minimum of $\phi(\log x, \log y)$ attained at a slope of $\beta = 1$ (i.e., 45°) and $r^2 = 1$.

To make our notation a bit easier again, let us write

$$r^2 \triangleq r^2(\log x, \log y)$$
$$\beta^2 \triangleq \hat{\beta}^2(\log x, \log y).$$

Rearranging terms gives us

$$B = A\beta^2$$
$$C^2 = r^2 A^2 \beta^2$$
$$C = \text{sign}(\beta) A |r\beta|$$

so that now we can write

$$\begin{aligned} \text{var}(\log(x/y)) &= A + B - 2C \\ &= A + A\beta^2 - 2\text{sign}(\beta) A |r\beta| \\ &= \text{var}(\log x)(1 + \beta^2 - 2\text{sign}(\beta)|r\beta|). \end{aligned} \tag{4}$$

Thus we have factored logratio variance into two terms[2]

- $\text{var}(\log x)$, which is about the magnitude of variability at play and has nothing to do with $\log y$
- $1+\beta^2-2\text{sign}(\beta)|r\beta|$, which is made up of slope and r^2 terms describing the relationship between $\log x$ and $\log y$ in a more interpretable way.

We use this second term to define

$$\phi(a, b) \triangleq 1 + \hat{\beta}^2(a, b) - 2\text{sign}(\hat{\beta}(a, b))\, |r(a, b)\, \hat{\beta}(a, b)|. \tag{5}$$

Figure 1 shows $\phi(\log x, \log y)$ as a function of the slope and coefficient of determination. When x and y are perfectly proportional, the slope of $\log y$ versus $\log x$ is 1, as is r^2 and, like $\text{var}(\log(x/y))$, $\phi(\log x, \log y) = 0$. However, unlike $\text{var}(\log(x/y))$, $\phi(\log x, \log y)$ does have an interpretable scale, e.g., values of $\phi(\log x, \log y) < 0.05$ indicate that x and y are highly proportional. Of course, qualifiers like "highly" are subjective and are used with similar intent as when applied to other statistical quantities, such as "significance". Speaking of which...

[2]At first blush, it might seem that we are treating x and y asymmetrically in this factorisation, but note that $\text{var}(\log(x/y)) = \text{var}(\log(y/x)) = \text{var}(\log y)(1 + 1/\beta^2 - 2\text{sign}(1/\beta)|r/\beta|)$.

2.4 An hypothesis testing alternative

$\phi()$ provides a measure of "goodness of fit to proportionality between x and y". However, similar to our companion paper, we have also considered assessing proportionality by testing the hypothesis that the slope of $\log y$ versus $\log x$ is 1. We will show results later to explain why we prefer the "goodness of fit" approach, at least for the molecular bioscience data we have investigated so far. Here, we explain the statistical machinery behind that test.

To ensure that the residuals involved in this hypothesis test remained in the simplex, we used the centered logratio transformation of the compositional data rather than its logarithm. To state that clearly, we have to abandon our simple notation for a moment and write the i^{th}sample of a D-part composition as $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$. Hypothesis tests about the slope of the first and second components use their centred logratio transformation, i.e., $\text{clr}_1(\mathbf{x}_i)$ and $\text{clr}_2(\mathbf{x}_i)$. Warton et al. (2006) describe a test of isometry, i.e., the hypothesis that the slope of the relationship between two random variables is equal to 1. This involves testing whether the *difference* of the two variables is uncorrelated to their *sum*, or in other words: if the data were rotated by 45°, would the subsequent values be uncorrelated?

3 Finding proportionality in yeast gene expression

In this section of the paper we demonstrate how $\phi(\text{clr}(x_i), \text{clr}(x_j))$ can be used to explore association in the sort of high-dimensional compositions commonly found in modern bioscience.

To provide compelling evidence that relative information requires special analysis and interpretation, we need to show how its naïve treatment can lead to very different conclusions from those obtained from measurements of absolute abundance. As we remarked in the beginning, relative abundance data is common in molecular bioscience; however, only recently have meaningful data on absolute abundance become available. We analysed data obtained by Marguerat et al. (2012) in a novel experiment to estimate the absolute abundance (i.e., copies per cell) of fission yeast messenger RNA (mRNA) over time, in the absence of a key nutrient.

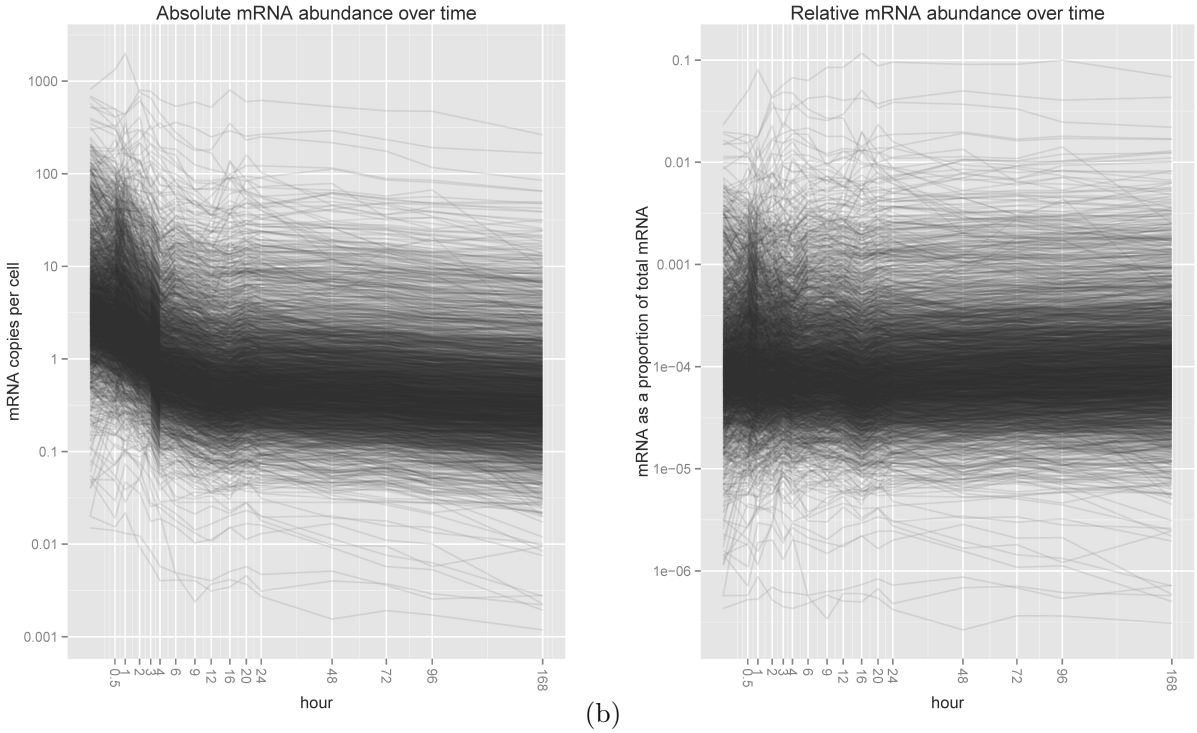
Our analyses are based on data from the RNA sequencing and microarray parts of the experiment which we combined to estimate the expression levels of over 7000 yeast mRNAs. We omitted records with zero or missing expression levels to obtain our primary dataset: 16 observations of the absolute abundance of 3031 different yeast mRNAs[3]. Figure 2 shows both the absolute and relative abundances of these mRNAs over time.

3.1 Correlations between relative abundances tell us absolutely nothing

Clearly, the absolute levels of gene expression are strongly positively correlated (i.e., changing in the same direction) over the time course; this is confirmed in histograms of correlation coefficients in the top margins of Figure 3. Figure 3(a) shows that the correlations between relative abundances (the marginal histogram on the right) give us no idea of the correlations between absolute abundances. Figure 3(b) shows the correspondence between values of $\phi()$ and the correlations of absolute abundances: very few strongly positively correlated pairs of mRNAs are also strongly proportional.

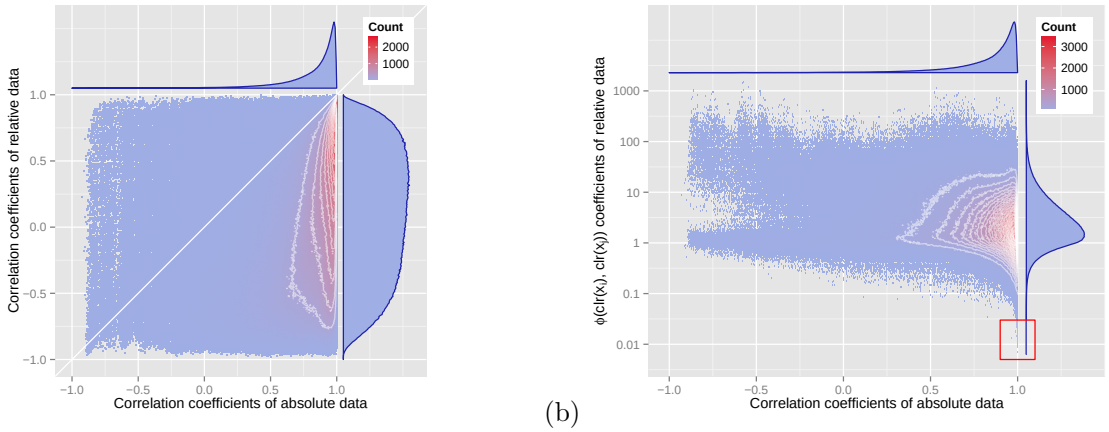
In the absence of any other information or assumptions, correlations between relative values tell us nothing about relationships between the absolute values from which they were derived. We stress *in the absence of any other information or assumptions* to highlight an assumption that underpins many gene expression studies: that the total level of gene expression (i.e., absolute abundance of all kinds of mRNA) remains fairly constant across all experimental conditions. If this assumption holds, the relative abundance of each kind of mRNA will be proportional to its absolute abundance, and analyses of correlation or "differential expression" of the relative values have clear interpretations.

[3]Eyebrows may be raised at our omission of over half the measured variables in this experiment. Obviously, this has no impact on the *absolute* abundances of the remaining variables. It does, however, significantly affect the apparent *proportions* of the remaining mRNAs and, if naïvely analysed, will have a similarly significant effect on the conclusions. The conclusions drawn from compositional data analysis are not affected by working with such subcompositions, a fact we will emphasise in presenting this work to bioscientists and others new to CoDA.



(a) (b)

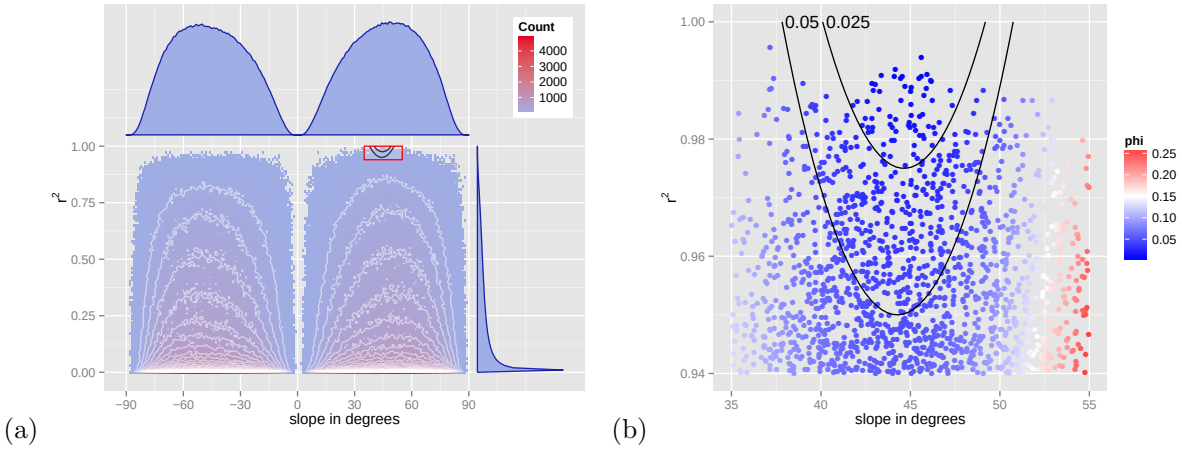
Figure 2: (a) Absolute and (b) relative abundances of 3031 yeast messenger RNAs over the 16-point time course from Marguerat et al. (2012). The y-axes of both plots are scaled logarithmically and the x-axes are on a square-root scale so that all the data can be clearly seen. Each grey line corresponds to the expression levels of a particular mRNA.



(a) (b)

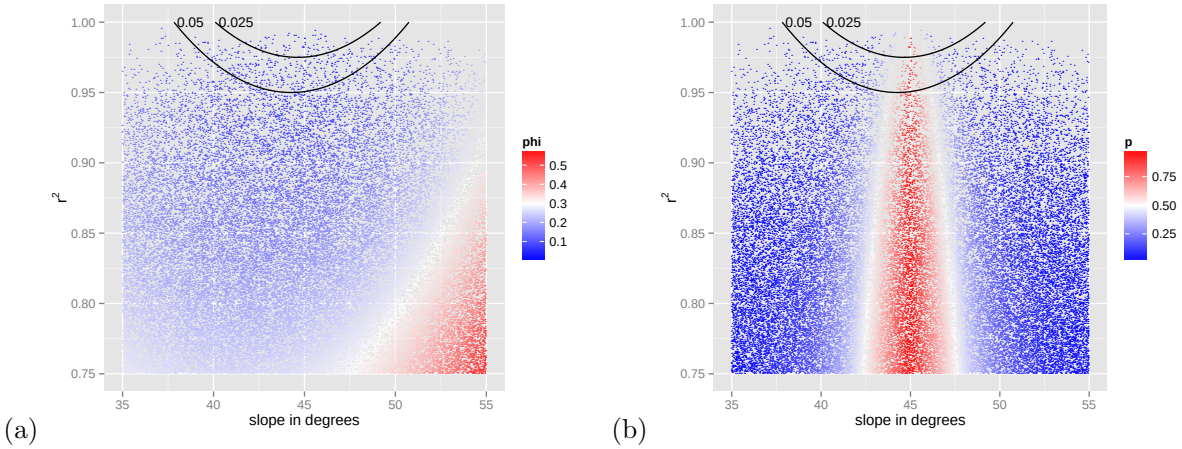
Figure 3: (a) a 2D histogram of the correlation coefficient observed for the relative abundances of a given pair of mRNAs, against the correlation coefficient observed for the absolute abundances of that same pair, over all pairs. White contour lines are shown at intervals of 100 counts. The top marginal histogram shows that the absolute abundances of most pairs are very strongly correlated. The right marginal histogram shows "the negative bias difficulty" (Aitchison, 1986, Section 3.3) of closure on correlation—here, correlations between relative abundances bear no relationship to the corresponding correlations between absolute abundances.
(b) a 2D histogram of $\phi(\mathrm{clr}(x_i), \mathrm{clr}(x_j))$ for the relative abundances of a given pair (i, j) of mRNAs, against the correlation coefficient observed for the absolute abundances of that same pair, over all pairs. White contour lines are again shown at intervals of 100 counts and the top marginal histogram is the same as in the left-hand figure. The few mRNA pairs that are strongly proportional (within the red rectangle) are also strongly positively correlated. However, the converse is not true: strong positive correlation between mRNAs does not imply that they are strongly proportional.

ISBN: 978-3-200-03103-6



(a) (b)

Figure 4: (a) The bivariate distribution of slope and r^2 values observed in all $3031 \times 3030/2 \approx 4.5$ million pairs of mRNA relative abundances in our time course. The marginal histograms show the distribution of slope values (top) and r^2 values (right). White contour lines are spaced at intervals of 100.
(b) A zoomed in view of the area inside the red rectangle in (a). The black lines show the 0.05 and 0.025 contours of $\phi(\mathrm{clr}(x_i), \mathrm{clr}(x_j))$ and points are coloured according to that statistic.



(a) (b)

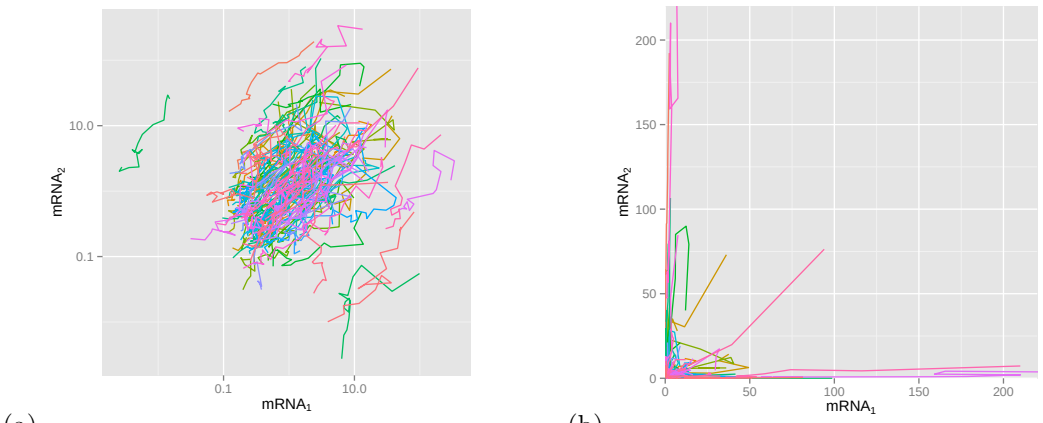
Figure 5: (a) A view of slope and r^2 values encompassing more data than Figure 4(b), again with points coloured by$\phi(\mathrm{clr}(x_i), \mathrm{clr}(x_j))$.
(b) The same points as (a) coloured this time by the p-value of the slope test of isometry.

Our understanding is that the assumption of constant gene expression is often implicit and seldom tested; the revisitation of this assumption by Lovén et al. (2012) should raise alarm bells about the inferences drawn from many gene expression studies.

As correlation is not an appropriate measure of association for data carrying only relative information (Lovell et al., 2013), correlation-based visualisations of relative abundance data (e.g., as described by Eisen et al. (1998)) should be abandoned in favour of visualisations based on proportionality, a topic we shall return to in Section 3.5.

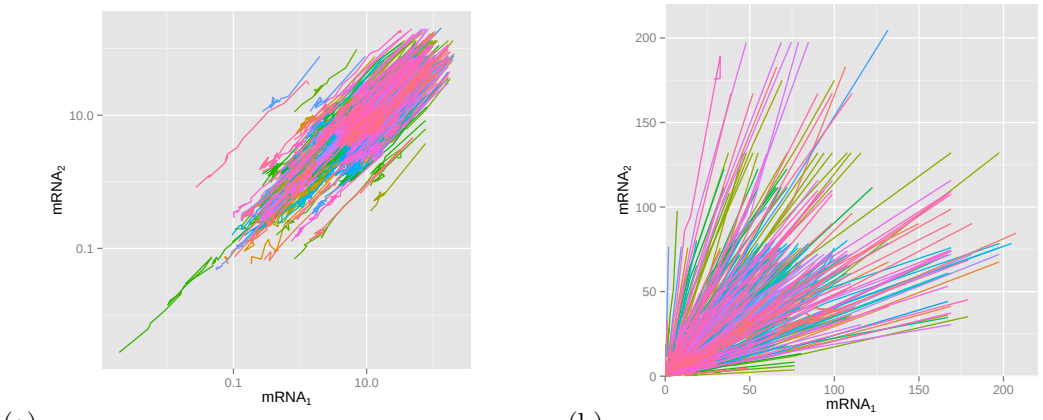
3.2 Strongly proportional components are also strongly correlated

Figure 3(b) plots the bivariate distribution of $\phi(\mathrm{clr}(x_i), \mathrm{clr}(x_j))$ and the correlation between the absolute abundances of mRNAs i and j over the time course. Here, the value of $\phi()$ between relative values *does* tell us something about relationships between the absolute values from which they were derived: mRNA pairs whose relative values are strongly proportions (with $\phi() < 0.05$, say) have strongly positively correlated absolute values ($r > 0.95$) as highlighted in the red rectangle in Figure 3(b).



(a) (b)

Figure 6: Absolute expression levels of the 136 pairs of mRNAs with slope test p-values> 0.9999 plotted (a) on a log-log scale and (b) on the natural scale.



(a) (b)

Figure 7: Absolute expression levels of the 424 pairs of mRNAs with $\phi(\mathrm{clr}(x_i), \mathrm{clr}(x_j)) < 0.05$ plotted (a) on a log-log scale and (b) on the natural scale.

3.3 Relatively few pairs of genes fit a proportional expression model here

Our "goodness of fit to proportionality" statistic, $\phi()$, comprises slope and r^2 terms describing the relationship between two components. Figure 4(a) plots the distribution of slope and r^2 values observed in all 4.5 million pairs of mRNA relative abundances in our time course. A tiny fraction of this number have r^2 and slopes that are both near 1 (the area within the red rectangle in Figure 4(a)). Figure 4(b) zooms in on these data and plots the slope and r^2 values for each pair.

3.4 Many more genes fail to reject the hypothesis of proportional expression

When we zoom out from the slope and r^2 values a little (Figure 5) we can start to investigate differences between our "goodness of fit to proportionality" statistic, $\phi()$, and the p-value underpinning the hypothesis test of isometry (Warton et al., 2006, Section V(1)). While $\phi()$ demands both slope and r^2 values be near 1 for a good fit to a proportional model, the hypothesis test assess the probability of obtaining a value of the test statistic more extreme than observed with our actual data, assuming the slope really is 1. (In this case, the test statistic is based on correlation of residuals to the fit of a line of slope 1.) If our p-value is small, we reject the null hypothesis that the slope is 1. If our p-value is not small, we fail to reject this null hypothesis. Roughly speaking, the hypothesis test focuses on the extent to which the data we saw for a given pair of mRNAs is *systematically* (as opposed to *randomly*) different from a pair that are proportional.

Statistical significance does not reflect the *strength* of proportional relationship. This becomes

clear if we plot all 136 pairs of mRNAs with slope test p-values > 0.9999 (6(a)) and contrast that with the 424 pairs of mRNAs with $\phi(\text{clr}(x_i), \text{clr}(x_j)) < 0.05$. The latter corresponds better to our intuitive notions of strong proportionality.

3.5 Visualising proportionality in high-dimensional compositions

$\phi()$ can be used to help visualise proportionality relationships in relative abundance data in much the same way that correlation can be used to help visualise relationships in absolute abundance data. To demonstrate this, we lay out a graph of the proportionality relationships between the 424 pairs of mRNAs with $\phi(\text{clr}(x_i), \text{clr}(x_j)) < 0.05$ in Figure 8. Each of the 217 unique nodes represents a different kind of mRNA. The 424 edges connect pairs of nodes that have $\phi(\text{clr}(x_i), \text{clr}(x_j)) < 0.05$ and the thickness of edge ij is inversely proportional to $\phi(\text{clr}(x_i), \text{clr}(x_j))$. This visualisation reveals sets of mRNAs whose expressions are strongly proportional. The largest of these involves 96 mRNAs, all of whose pairwise values of $\phi()$ can be visualised using the clustered heatmap approach commonly applied to molecular bioscience data (Figure 9).

The hierarchical clustering used to rearrange the matrix of $\phi()$ values in Figure 9 can help analysts explore clusters of components that exhibit distinct patterns. To demonstrate this in action, Figure 10 takes the 66 pairs of mRNAs with $\phi(\text{clr}(x_i), \text{clr}(x_j)) < 0.025$, and cuts the hierarchy into six groups whose absolute and relative abundances over time are shown in Figure 11.

4 In conclusion

Molecular bioscience provides many data analysis challenges. Making sound inference from small numbers of observations with large numbers of variables has been an obvious challenge from the outset. Less obvious, but perhaps more fundamental, is the challenge of analysing and interpreting measurements that carry only relative information. Compositional data analysis has much to contribute towards this and, in broadening its application to the biosciences, it is stimulated to address challenges not found in its traditional domains of application.

This paper sets out new ideas in compositional data analysis so that its principles can be more readily applied to molecular bioscience. In particular, we have shown how CoDA's logratio variance can be made more interpretable by drawing on the standardised major axis approach to bivariate line fitting. This, in turn, allows us to propose a measure of association $\phi()$ that can be confidently applied to relative abundance data instead of correlation which, while commonplace, is manifestly inappropriate for such data. We have shown that $\phi()$, as a "goodness of fit to proportionality" statistic, offers a more intuitive approach than hypothesis testing, at least in its application to exploring association in high-dimensional yeast gene expression data.

Finally, we have demonstrated how $\phi()$ can be straightforwardly employed in some multivariate analysis and visualisation strategies commonly employed in molecular bioscience, including network inference, heatmaps and hierarchical clustering.

Certainly, there is much more to be done to increase the benefits that compositional data analysis can bring to molecular bioscience. Our hope is that the ideas in this paper will be a useful step in that direction.

5 Acknowledgments

We thank Samuel Marguerat and Jürg Bähler for their comments, suggestions and data. This research has been supported by the Spanish Ministry of Education, Culture and Sports under a Salvador de Madariaga grant (Ref. PR2011-0290); by the Spanish Ministry of Economy and Competitiveness 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 project Ref: 2009SGR424. We also gratefully acknowledge the developers of the software and analysis packages that profoundly enabled this research, in particular The R Core Team (2012), Wickham (2009), Xie (2012), van den Boogaart et al. (2012), RStudio Incorporated (2013) and Bastian et al. (2009).

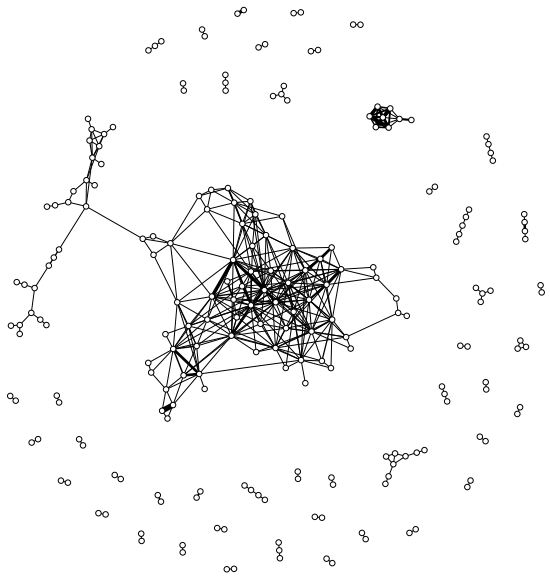


Figure 8: A graph of the proportionality relationships between the 424 pairs of mRNAs with $\phi(\text{clr}(x_i), \text{clr}(x_j)) < 0.05$.

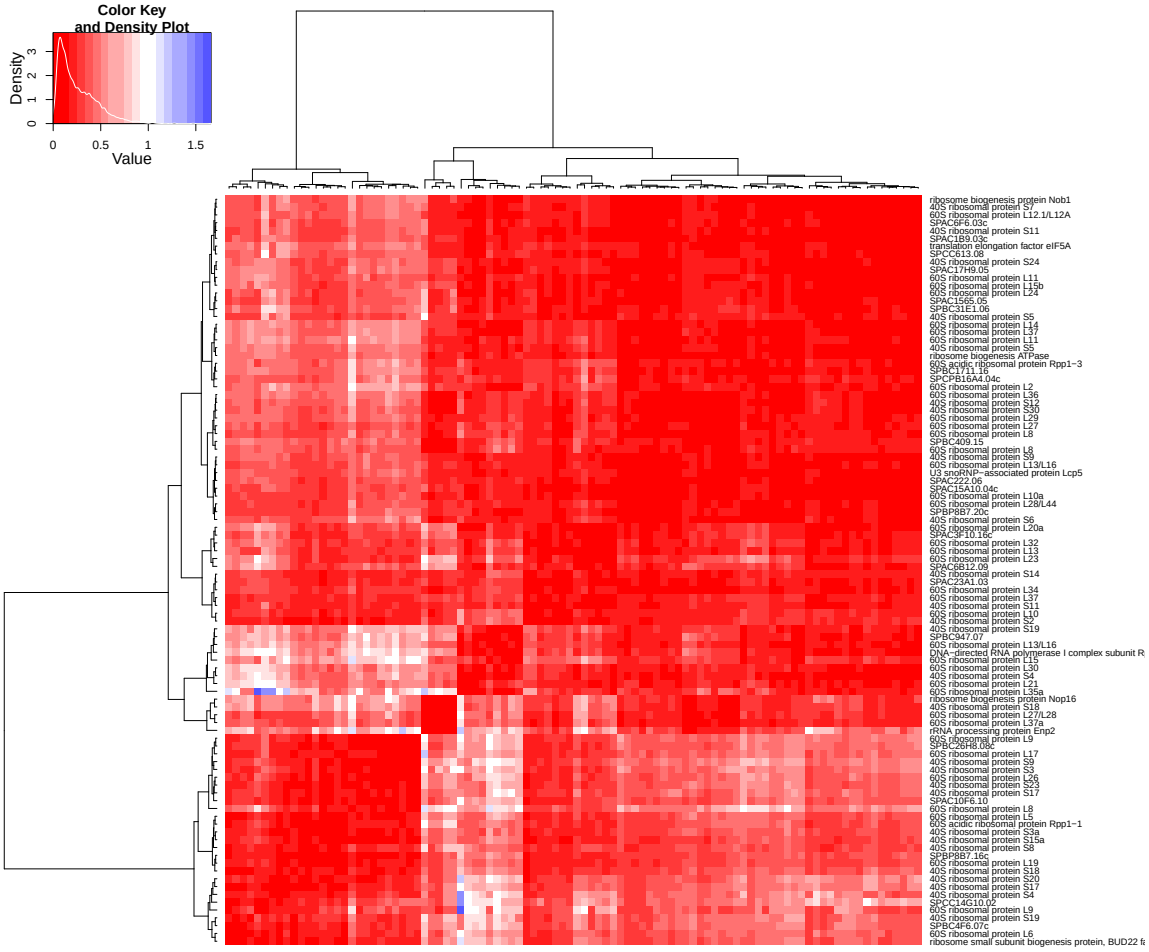


Figure 9: Heatmap visualisation of the 96 mRNA cluster seen in Figure 8.

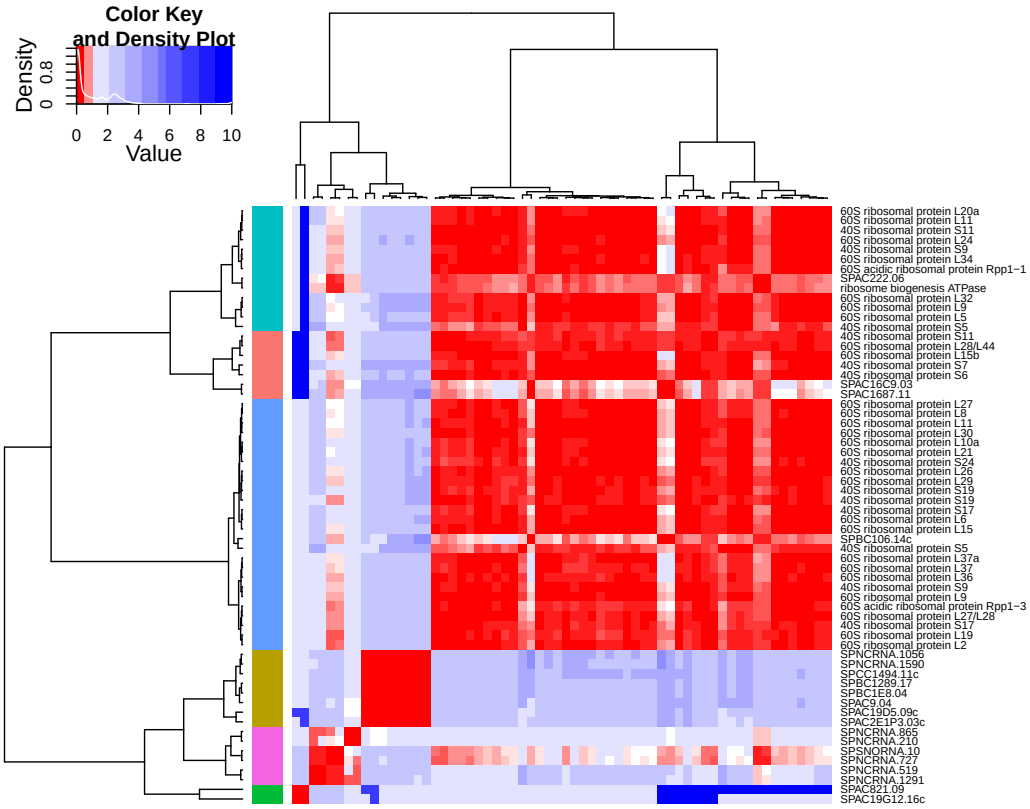


Figure 10: Heatmap visualisation of the 66 pairs of mRNAs with $\phi(\mathrm{clr}(x_i), \mathrm{clr}(x_j)) < 0.025$. The hierarchical clustering of these components is cut into six colour-coded groups, shown at the left edge of the heatmap.

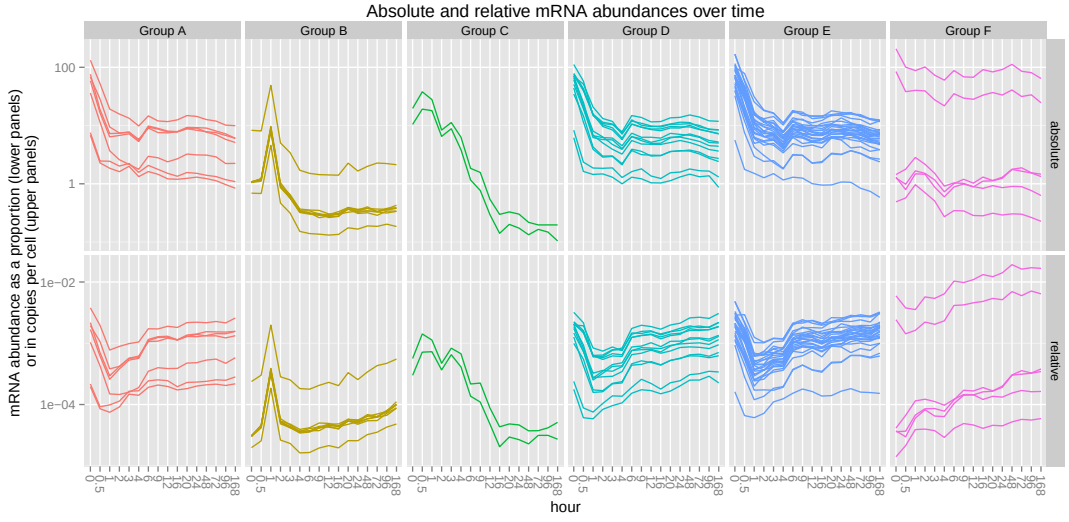


Figure 11: Absolute and relative abundances of the 66 pairs of mRNAs clustered into six groups in Figure 10. The line colours correspond to the colour-coding of groups in Figure 10.

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman & Hall, Ltd.

Bansal, M., V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo (2007, February). How to infer gene networks from expression profiles. *Molecular Systems Biology 3*, 78. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1828749/.

Bastian, M., S. Heymann, and M. Jacomy (2009). Gephi: An open source software for exploring and manipulating networks. http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154.

Buccianti, A., G. Mateu-Figueras, V. Pawlowsky-Glahn, and Editors (2006, November). *Compositional Data Analysis in the Geosciences: From Theory to Practice - Special Publication no 264* (illustrated edition ed.). Geological Society of London.

Eisen, M. B., P. T. Spellman, P. O. Brown, and D. Botstein (1998, December). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences 95*(25), 14863–14868. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC24541/.

Faust, K., J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, and C. Huttenhower (2012, July). Microbial co-occurrence relationships in the human microbiome. *PLoS Comput Biol 8*(7), e1002606. http://dx.doi.org/10.1371/journal.pcbi.1002606.

Friedman, J. and E. J. Alm (2012, September). Inferring correlation networks from genomic survey data. *PLoS Comput Biol 8*(9), e1002687. http://dx.doi.org/10.1371/journal.pcbi.1002687.

Lovell, D., W. Müller, J. Taylor, A. Zwart, and C. Helliwell (2011). Proportions, percentages, PPM: do the molecular biosciences treat compositional data right? In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional Data Analysis: Theory and Applications*, pp. 191–207. Chichester, UK: John Wiley & Sons, Ltd. http://onlinelibrary.wiley.com/doi/10.1002/9781119976462.ch14/summary.

Lovell, D., V. Pawlowsky-Glahn, and J. J. Egozcue (2013, March). Don't correlate proportions! http://www.slideshare.net/AustralianBioinformatics/dont-correlate-proportions.

Lovén, J., D. A. Orlando, A. A. Sigova, C. Y. Lin, P. B. Rahl, C. B. Burge, D. L. Levens, T. I. Lee, and R. A. Young (2012, October). Revisiting global gene expression analysis. *Cell 151*(3), 476–482. http://www.cell.com/abstract/S0092-8674(12)01226-3.

Marguerat, S., A. Schmidt, S. Codlin, W. Chen, R. Aebersold, and J. Bähler (2012, October). Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell 151*(3), 671–683. http://www.cell.com/abstract/S0092-8674(12)01126-9.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org/.

RStudio, Inc. (2013). *RStudio: Integrated development environment for R*. http://www.rstudio.org.

van den Boogaart, K. G., R. Tolosana, and M. Bren (2012). *compositions: Compositional Data Analysis*. http://CRAN.R-project.org/package=compositions.

Warton, D. I., I. J. Wright, D. S. Falster, and M. Westoby (2006). Bivariate line-fitting methods for allometry. *Biological Reviews 81*(2), 259–291. http://onlinelibrary.wiley.com/doi/10.1017/S1464793106007007/abstract.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer New York. http://had.co.nz/ggplot2/book.

Xie, Y. (2012). *knitr: A general-purpose package for dynamic report generation in R*. http://CRAN.R-project.org/package=knitr.

Vulnerability model for a nuclear power plant containment building.

A. MUSOLAS[1], J. J. EGOZCUE[1], and M. CRUSELLS[2]

[1]Dept. of Applied Mathematics III, [2]Dept. of Construction Engineering,
U. Politècnica de Catalunya, Barcelona, antoni.musolas@estudiant.upc.edu

Abstract

When supervising a nuclear power plant, the containment building is crucial. Its functions are guaranteeing structural integrity and avoiding leaks in case of accident. Both events are considered of high risk. Once a given overpressure is registered inside the containment building, three possible outputs are considered: serviceability, breakdown, and collapse. The current aim is the study of vulnerability. The vulnerability of the containment building under overpressure is described by the conditional probability of the three mentioned outputs.

The study consists of three steps: (a) modelling the containment building using the Finite Element Method; (b) given an overpressure, simulating uncertain parameters related to material constitutive equations in order to obtain the corresponding output; (c) performing a simplicial regression to get a meaningful vulnerability model. The simulation provides proportions of outputs under the overpressure conditions. The vulnerability model can be obtained by a simplicial regression of those proportions, as a response variable, on the overpressure, as explanatory variable.

Some difficulties appear in step (c). For any given overpressure, some outputs are very improbable, producing zeros in the corresponding proportion. In order to reduce the presence of zero proportions, importance re-sampling is applied in the simulation. Importance re-sampling allows the simulation of favourable conditions to get the desired output. Then, importance of simulated data is taken into account when fitting the vulnerability model in step (c). The obtained vulnerability model is similar to previous results in nuclear power plant safety analysis.

1 Introduction to the containment building model

When supervising a nuclear power plant, the containment building is crucial. Its functions are guaranteeing structural integrity and avoiding leaks in case of accident. Both events are considered of high risk. A containment building in a nuclear power plant is typically a reinforced or prestressed concrete structure enclosing a nuclear reactor. Moreover, in the inner part of the containment, there is a metallic shell, the liner; its function is to avoid the leakage in case of accident. This study is focused on the analysis of the structural integrity of the containment building, so the liner has not been taken into account in the model.

The containment building is designed to contain the escape of radiation in any emergency to a maximum pressure relative pressure in the range of 0.4 to 1.4 MPa. The containment is the fourth and final barrier to radioactive release after the fuel ceramic, the metal fuel cladding, and the reactor pressure boundary. From now on, the pressure will be definite as relative, which is the absolute pressure minus the one in the atmosphere (0.1 MPa).

Containment systems for nuclear power reactors are distinguished by size, shape, material, and reactor coolant state. In this analysis, a three-loop Pressurized Water Reactor (PWR) is considered. For this type of reactor, the containment also encloses the steam generators, the pressurizer, and the entire reactor coolant system. Early designs by Siemens, Westinghouse, and Combustion Engineering had a mostly can–like shape built with reinforced or prestressed concrete. However, depending on the material used, the most apparently logical design is a can-like topped by a half-spherical top, which is usually the best structure for simply containing a large pressure. In this case, the model will be a prestressed cylindrical building with a half-spherical top. The most important dimensions of the building are shown in Table 1 and in Figure 1.

Dimension	Value
Inner diameter (m)	40
Total inner height (m)	63.4
Cylinder inner height (m)	43.4
Foundation thickness (m)	3
Cylinder thickness (m)	1.15
Dome thickness in the top (m)	0.95

Table 1: Most important dimensions of the building.

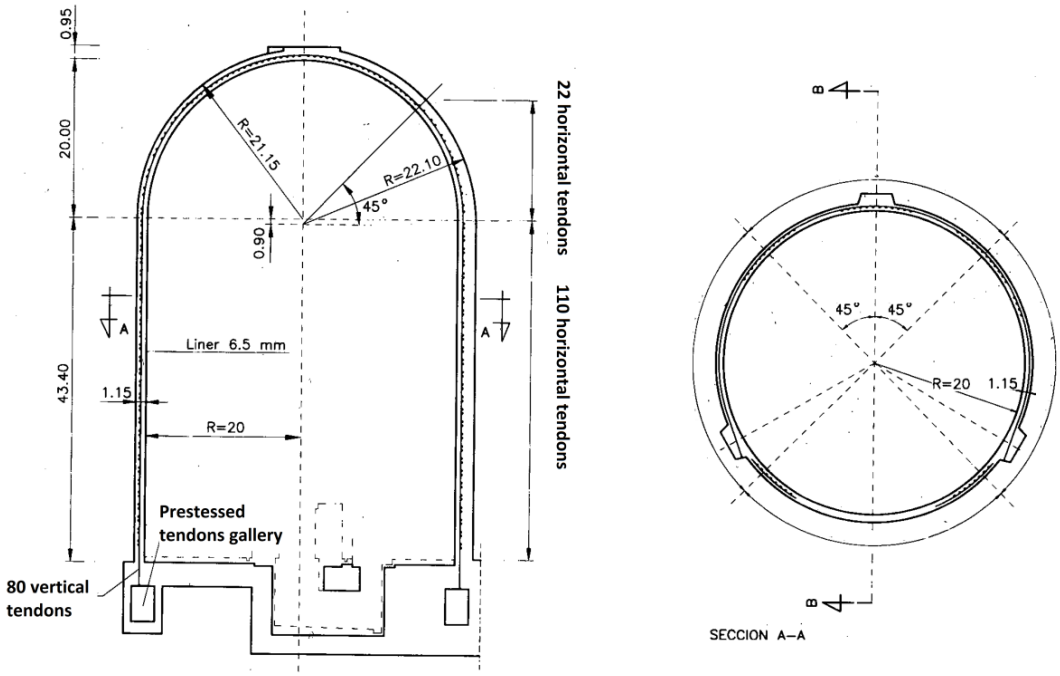


Figure 1. Scheme of the building. Cervera et al. (1985).

According to these schemes, it is possible to draw a geometrical model in a finite element software such as Abaqus®. The additional inputs required in the model are the parameters related to the material properties. The most important parameters of the building materials are shown in Table 2.

Concrete parameters	Value
Elastic modulus (GPa)	30
Compressive strength (MPa)	39
Tensile strength (MPa)	3
Density (kN/m^3)	25
Poisson coefficient (-)	0.2

Reinforcing beam parameters	Values
Elastic modulus (GPa)	220
Elastic yield strength (MPa)	470
Ultimate tensile strength (MPa)	610

Prestressing tendon parameters	Values
Vertical tendons prestressing force (kN)	4700
Horizontal tendons prestressing force (kN)	5400
Dome tendons prestressing force (kN)	5500

Table 2: Most important materials properties.

Once all the above is properly introduced, it is required to apply the constraints and boundary conditions that consist basically of fixing the foundation. Finally, the last input is the external load which is an inner pressure. Figure 2 shows that the deformation when the building is faced up to an overpressure of 0.7 MPa: the building becomes shorter but wider. Figure 2 (right) colours in red the elements that have reached plasticity.

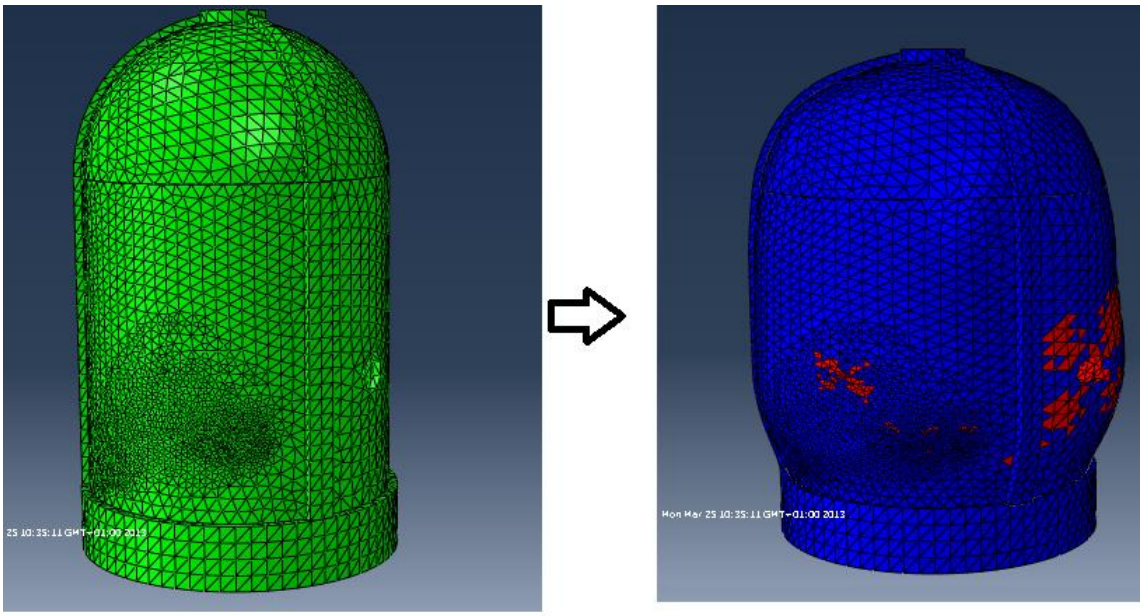


Figure 2. Mesh of the finite element model of the building (left). Deformation caused by a 0.7 MPa inner pressure, only red elements have reached plasticity (right).

When the containment building faces an overpressure, the response can be classified into three possible states: serviceability, breakdown, and collapse. These states will be defined below. The vulnerability of the containment building under overpressure is described by the conditional probability of the three mentioned outputs. The vulnerability study consists of three steps: (a) modelling the containment building using a finite element software as shown above; (b) given an overpressure, simulating uncertain parameters related to material constitutive equations in order to obtain the corresponding output; (c) performing a simplicial regression to get a meaningful

vulnerability model. The simulation provides proportions of outputs under the over-pressure conditions. The vulnerability model can be obtained by a simplicial regression of those proportions, as a response variable, on the overpressure, as explanatory variable.

2 Random variables

One of the most common problems in Civil Engineering is that material properties are usually unknown. It is fairly difficult to ascertain their values even with an extensive test campaign. In order to capture this uncertainty in the material properties, five variables have been taken as random and four other variables have been taken directly depending on the random ones. Applying this directly dependence, it is possible to account in a simple way some relations between variables. For instance, the elasticity modulus of concrete is a function of its compressive strength. However, it has been supposed that the material is uniform in the whole building. As a consequence, the value of material parameters is the same in all its parts, in each computation.

The random parameters and their distribution are shown below:
1.- Concrete elastic modulus (Em). LogNormal distribution, E[Em]=41100 MPa, Var[Em]=16892100 MPa2, i.e. logarithmic mean and variance are 6.0638, 0.05743, respectively.
2.- Steel elastic yield strength (fy). LogNormal distribution, E[fy]=430 MPa, Var[fy]=576 MPa2, i.e. logarithmic mean and variance are 6.0638, 0.05743, respectively.
3.- Vertical cables prestressing force. Logistic-Normal distribution on the interval (0,610). The logistic parameters have mean 1.30 and standard deviation 0.43. The median of this distribution is 610•exp(1.30)/(1+exp(1.30))=4790 kN.
4.- Horizontal cables prestressing force. Logistic-Normal distribution on the interval (0,610). The logistic parameters have mean 2.22 and standard deviation 0.80. The median of this distribution is 610•exp(2.22)/(1+exp(2.22))=5500 kN.
5.- Dome cables prestressing force. Logistic-Normal distribution on the interval (0,610). The logistic parameters have mean 2.36 and standard deviation 0.71. The median of this distribution is 610•exp(2.36)/(1+exp(2.36))=5570 kN.

The parameters of the mentioned distributions have been based on a huge data set from Aguado et al. (1991) and Barbat et al. (1995).

The variables that have a direct dependence on the random parameters are:
1.- Concrete compressive strength. fc=(Em-1550)/697
2.- Concrete tensile strength. ft=0.30·(fc-8)^(2/3)
4.- Steel elastic modulus. Es=fy/0.00214
5.- Steel ultimate strength. fu=(5500/4200)·fy

These relations have been taken from EHE-08 and Aguado et al. (1991). Other parameters such as the density of the concrete or the Poisson coefficient have been taken as fixed values since their variance is actually fairly unnoticeable.

3 Damage levels

Once a given overpressure is registered inside the containment building, three possible outputs are considered: serviceability, breakdown, and collapse. In order to know in which final state the structure is, the failure criterion has to be defined. To do it simply, the largest value of the maximum principal tensile strain (S11) in the whole building has been extracted in each computation. Two thresholds in S11 define the three possible outputs. For S11 less than 0.3 mm/m serviceability is assumed. From S11 equal to 0.3 mm/m to 2 mm/m, the output is considered breakdown. Finally, for S11 greater than or equal to 2 mm/m, the output is considered collapse. The 0.3 mm/m threshold typically indicates that concrete has already cracked and the resistance is then controlled by the

reinforcing beams. The second one, between breakdown and collapse, has been fixed in 2 mm/m which is the threshold that indicates that the reinforcing beams have reached plasticity.

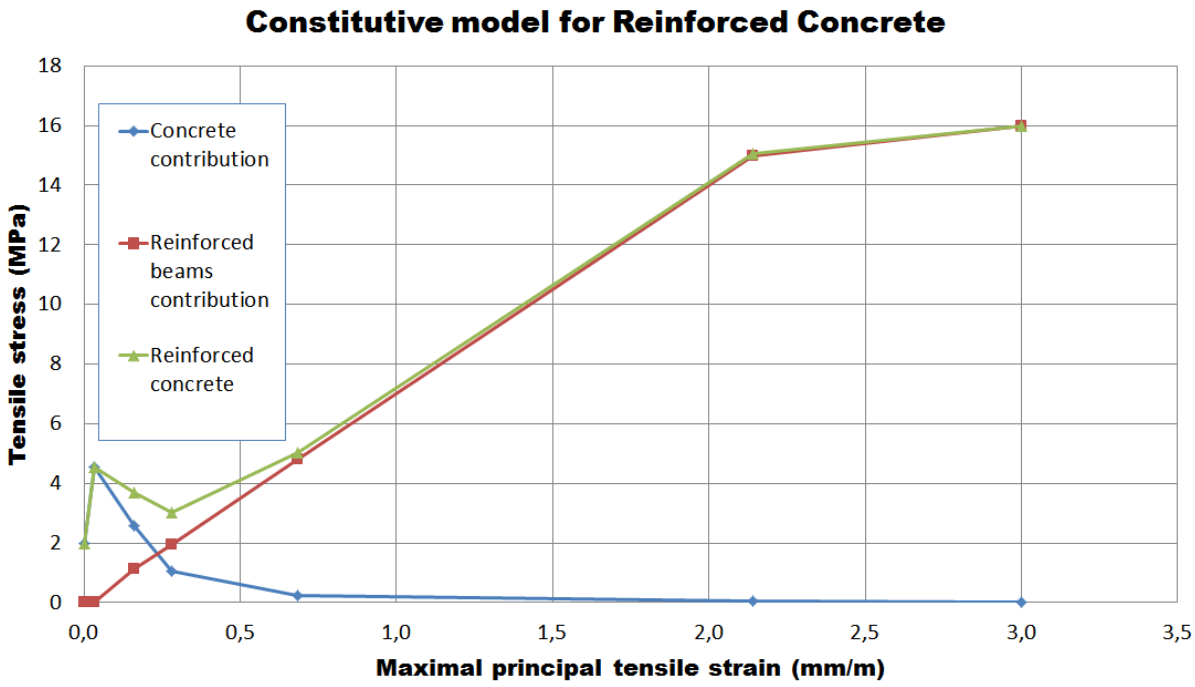


Figure 3: Nonlinear tensile constitutive behaviour.

These thresholds clearly split the three possible final scenarios. Serviceability is related to no damage in the building because the concrete has not cracked yet. So, the nuclear power plant can restart its operations without any concern about its structural integrity. In the case of breakdown, the building has been slightly damaged because in some parts the concrete has started to crack. The nuclear power plant has to stop its activity and seal several cracks before restarting the activity. Finally, in the case of collapse (Figure 3), the building is completely damaged because even the reinforcing beams have reached plasticity and it has to be demolished and rebuilt.

4 Parameter simulation and vulnerability model

Taking all the distributions of random parameters into account, the procedure to simulate the parameters is reproduced as follows. First, the overpressure is fixed to a certain value $p*$. Second, the five random parameters are simulated as independent random variables. Once the random variables are simulated, the directly dependent parameters are obtained using the corresponding expressions. Then, for the overpressure value $p*$ and for each set of simulated parameters, the strains are computed using the Finite Elements Method. From the computed strains, the output (serviceability, breakdown, and collapse) is determined in each simulation. Given a $p*$, the proportion of computations in which each possible output appears is obtained.

To do that, a multinomial random variable Z is defined. Its parameters are functions of the overpressure and any realization of Z is a vector containing a single 1 in the three components. The indexes refer to serviceability, breakdown, and collapse, respectively. For instance, if the vector $\boldsymbol{q}_i$, which contains the inputs in the computation i, gives an output of collapse, the realization of Z would be: (0, 0, 1).

So, the desired probabilities are:

$$\begin{aligned} P[Z=(1,0,0)|p^*] &= x_1(p^*), \\ P[Z=(0,1,0)|p^*] &= x_2(p^*), \\ P[Z=(0,0,1)|p^*] &= x_3(p^*). \end{aligned} \quad (1)$$

These components are the expectation of Z related to the overpressure p^* and so, they can be computed using the Monte Carlo integration method.

$$\boldsymbol{x}(p^*) = [x_1(p^*), x_2(p^*), x_3(p^*)] = E[Z|p^*] = \int \boldsymbol{z}(\boldsymbol{q}|p^*) f_{\boldsymbol{q}}(\boldsymbol{q}) d\boldsymbol{q} \approx \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{z}(\boldsymbol{q}_i|p^*). \quad (2)$$

The main problem is that for any given overpressure, some outputs are very improbable, producing zeros in the corresponding proportions. This is the case of, for instance, the $x_3(p^*)$ with $p^* = 0.4$ MPa. In this case, the overpressure is so low that it is fairly impossible to get a simulation that ends with collapse. In order to reduce the presence of zero proportions, importance re-sampling (Hammersley, J. M. and D. C. Handscomb, 1964; Robert, C. and G. Casella, 2004) has been applied in the simulation.

The procedure is the following: the sampling is done in terms of a distribution, called sampling distribution with density function $\hat{f}_{\boldsymbol{q}}(\boldsymbol{q}_i)$, which allows desired parameter values likely appearing in the simulation. Then, the importance $f_{\boldsymbol{q}}(\boldsymbol{q}_i)/\hat{f}_{\boldsymbol{q}}(\boldsymbol{q}_i)$ is the rate of the model probability density function over the sampling probability density function. This importance ratio is stored in each simulation and the Monte Carlo procedure is modified as follows:

$$\boldsymbol{x}(p^*) = E[Z|p^*] = \int \boldsymbol{z}(\boldsymbol{q}|p^*) \hat{f}_{\boldsymbol{q}}(\boldsymbol{q}) \frac{f_{\boldsymbol{q}}(\boldsymbol{q})}{\hat{f}_{\boldsymbol{q}}(\boldsymbol{q})} d\boldsymbol{q} \approx \frac{1}{m}\sum_{i=1}^{m} \boldsymbol{z}(\boldsymbol{q}_i|p^*) \frac{f_{\boldsymbol{q}}(\boldsymbol{q}_i)}{\hat{f}_{\boldsymbol{q}}(\boldsymbol{q}_i)}. \quad (3)$$

Table 3 illustrates the methodology to obtain the vector $\boldsymbol{x}(p^*)$:

Simulated parameters	Serviceability	Breakdown	Collapse	Importance ratio
$\boldsymbol{q}_1$	1	0	0	1.51
$\boldsymbol{q}_2$	1	0	0	4.21
$\boldsymbol{q}_3$	0	0	1	0.04
...	...	...	...	...
$\boldsymbol{q}_{2000}$	0	1	0	0.53

Table 3: Scheme of importance re-sampling. Output data.

As mentioned previously, $\boldsymbol{q}_i$ includes all the inputs required by the software in the simulation i. The output is the final state of the building: serviceability, breakdown, or collapse and the importance is the rate of the model probability density function over the sampling probability density function $f_{\boldsymbol{q}}(\boldsymbol{q}_i)/\hat{f}_{\boldsymbol{q}}(\boldsymbol{q}_i)$. After $m = 2000$ the value of the three proportions of $\boldsymbol{x}(p^*)$ is obtained as a relative frequency. These proportions are denoted $[x_1(p^*), x_2(p^*), x_3(p^*)]$ and they are related to the overpressure p^*.

Since the re-sampling has been applied when simulating the inputs, the number of occurrences for each possible output has to be weighted by its importance. Therefore, in each computation, the value obtained is the importance rate in the corresponding component. The proportions of each component are also computed in terms of the sum of importances. Table 4 illustrates the procedure to obtain these components.

Simulated parameters	Serviceability	Breakdown	Collapse	Importance ratio
$\mathbf{q}_1$	1.51	0	0	1.51
$\mathbf{q}_2$	4.21	0	0	4.21
$\mathbf{q}_3$	0	0	0.04	0.04
...	...	...	...	...
$\mathbf{q}_{2000}$	0	0.53	0	0.53
Sum	1056.23	102.42	12.77	1171.42
Proportion $x(p^*)$	0.9017	0.0874	0.0109	

Table 4: Scheme of importance re-sampling. Modified output data.

$$x_1(p^*)=\frac{1056.23}{1171.42}=0.9017,\ x_2(p^*)=\frac{102.42}{1171.42}=0.0874,\ x_3(p^*)=\frac{12.77}{1171.42}=0.0109. \quad (4)$$

Until here, the probabilities of each possible output (serviceability, breakdown, and collapse) have been obtained for a given overpressure p^*. However, the vulnerability model is $\hat{\boldsymbol{x}}(p)$ for any overpressure.

The overpressure is characterized by its value in MegaPascals (MPa) and it is possible to discretize this value in thirteen numbers from 0.4 to 1.0 in a 0.05 step. Then, the data is described by the probabilities of each three final states conditioned to each 13 values of internal pressure. The problem can be seen as a linear least squares fitting. The data to be fitted are the composicional data points $\boldsymbol{x}(0.40), \boldsymbol{x}(0.45)....\boldsymbol{x}(1.00)$; each of them containing the probability of each three final possible states. For the sake of simplicity, this data points $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}....\boldsymbol{x}^{(13)}$. In this case, the explanatory variable is the overpressure p from 0.4 to 1.0 in MPa. Therefore, the vulnerability model $\hat{\boldsymbol{x}}(p)$ will be obtained for any overpressure.

Two problems are involved in this approach: the consistency of the model and the relative scale of the probability values. Since the vulnerability model is described by probabilities that must sum one, even small deviations in estimation in a probability value can result in an inconsistency of the model. Moreover, we are used to measure probabilities relatively; they are not a mere absolute value in the real numbers scale. These difficulties suggest approaching the problem by means of the simplex geometry of D parts (Aitchison, 1986; Pawlowsky-Glahn and Egozcue, 2001; Egozcue et al., 2003) that allows interpolating probability vectors in a consistent way and in an appropriate scale. The elements in the simplex (S^D) have D probability components but the dimension is D-1. It could be proved that, two operations similar to the addition and the multiplication exist and they form a vector space (Aitchison, 1986; Aitchison, 2002). Moreover, S^D is a D-1 dimensional Euclidian space (Pawlowsky-Glahn and Egozcue, 2001) if its own distance is added. This distance is called Aitchison distance as it was introduced by Aitchison (Aitchison, 1986).

Then, the data that has to be fitted is $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} .... \boldsymbol{x}^{(13)}$ and the predictor function in the simplex is:

$$\hat{\boldsymbol{x}}(p) = \boldsymbol{\beta}_0 \oplus (p \otimes \boldsymbol{\beta}_1). \quad (5)$$

In Equation 5, $\oplus$ and $\otimes$ are perturbation and powering in the simplex. The values of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are compositional parameters to be fitted in the regression.

This linear fitting can be reduced to a D-1 standard linear regression models that can be fitted using least squares techniques. The procedure consists of expressing the composition $\hat{\boldsymbol{x}}(p)$ into orthonormal coordinates using an ilr transformation (Egozcue et al., 2003). In this case, the operations $\oplus$ and $\otimes$ are reduced to the common + and ·, respectively. Then, S^D is considered equivalent to $\mathbb{R}^{D-1}$, when compositions are represented by orthonormal coordinates. These facts allow transforming the proportions $x_1(p), x_2(p), x_3(p)$ into balance-coordinates $b_1(p)$ and $b_2(p)$. A regression of these $b_1(p)$ and $b_2(p)$ on overpressure provides a linear vulnerability model in $\mathbb{R}^{D-1}$ (Egozcue et al., 2012).

An easy way to obtain orthonormal coordinates is producing a sequential binary partition (SBP) (Egozcue and Pawlowsky-Glahn, 2005). Table 5 shows the code of such SBP. In the first step breakdown and collapse are separated from seviceability as shown by the signs 1 and -1. The second and last step consists in separating breakdown from collapse.

order	Serviceability	Breakdown	Collapse
1	1	-1	-1
2	0	1	-1

Table 5: SBP code used to build up balance-coordinates using ilr.

The balance-coordinates corresponding to the SBP (Table 5) are:

$$b_1(p) = \sqrt{\frac{2}{3}} \cdot \log\left(\frac{x_1(p)}{\sqrt{x_2(p) \cdot x_3(p)}} \right); \; b_2(p) = \sqrt{\frac{1}{2}} \cdot \log\left(\frac{x_2(p)}{x_3(p)} \right). \quad (6)$$

Then, the data that has to be fitted in the regression model are $\boldsymbol{b}^{(1)}, \boldsymbol{b}^{(2)} .... \boldsymbol{b}^{(13)}$, where $\boldsymbol{b}^{(i)} = \left(b_1^{(i)}, b_2^{(i)}\right)$ is a vector in $\mathbb{R}^{D-1}$:

$$\boldsymbol{b}(p) = \boldsymbol{\beta}_0^* + (p \cdot \boldsymbol{\beta}_1^*), \quad (7)$$

In Equation 7, $\boldsymbol{\beta}_0^*$ and $\boldsymbol{\beta}_1^*$ are the coordinates of $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$, i.e. they are $\mathbb{R}^{D-1}$ vectors.

Then, this vector expression can be broken down into two pieces, one for each component:

$$\begin{aligned} b_1(p) &= \beta_{0,1}^* + (p \cdot \beta_{1,1}^*), \\ b_2(p) &= \beta_{0,2}^* + (p \cdot \beta_{1,2}^*). \end{aligned} \quad (8)$$

For each linear least squares fitting, the function that has to be minimized is:

$$SSE_1 = \sum_{i=1}^{13} \left| b_1(p) - b_1^{(i)} \right|^2 ,$$
$$SSE_2 = \sum_{i=1}^{13} \left| b_2(p) - b_2^{(i)} \right|^2 . \quad (9)$$

Finally, all above can be retransformed into the previous space (S^D) using ilr^{-1} for easy interpretation.

4 Results and conclusions

Mentioned above, the probabilities of each output -serviceability, breakdown, and collapse- in each overpressure situation have been obtained by proportion between importance rates. Table 6 shows the results obtained and the number of simulations carried out to obtain them.

Pressure Value (MPa)	Serviceability	Breakdown	Collapse	Number of simulations
0,40	0,9977145906	0,0022852621	0,0000001472	2384
0,45	0,9895539389	0,0104403679	0,0000056931	2257
0,50	0,9621359541	0,0378386536	0,0000253924	2250
0,55	0,9009824466	0,0987656199	0,0002519335	2188
0,60	0,7007071120	0,2983557187	0,0009371693	2240
0,65	0,3106313944	0,6825096513	0,0068589543	1780
0,70	0,0012657068	0,9791853629	0,0195489303	1794
0,75	0,0000000000	0,9666605764	0,0333394236	1477
0,80	0,0000000000	0,8539471646	0,1460528354	1852
0,85	0,0000000000	0,6888730503	0,3111269497	1375
0,90	0,0000000000	0,2135879397	0,7864120603	1345
0,95	0,0000000000	0,0099305357	0,9900694643	1191
1,00	0,0000000000	0,0000000000	1,0000000000	1165
				23298

Table 6: Final data obtained from a simulation.

Once the data above is computed, it is possible to obtain the balances trough the isometric-logratio transformation. However, some ratios are not possible since some components are zero. It is noticeable that the importance re-sampling has helped a lot in eliminating those zero entries. However, some extreme data such as the value of b_1 in the overpressure 0.70 and the value of b_2 in the overpressure 0.95 will not be used in the linear regression. The reason is that both come from data that is quite difficult to obtain in the simulation, even applying the importance re-sampling technique, and so, they contain a considerable error. In case both points would be used, they would have distorted the linearity of the regression.

Pressure Value (MPa)	b_1	b_2
0,40	8,9030	6,8235
0,45	6,7840	5,3133
0,50	5,6250	5,1666
0,55	4,2428	4,2224
0,60	3,0499	4,0752

0,65	1,2353	3,2528
0,70	~~-3,8328~~	2,7675
0,75	-	2,3809
0,80	-	1,2487
0,85	-	0,56205
0,90	-	-0,92167
0,95	-	~~-3,2542~~
1,00	-	-

Table 7: Logratio transformed data.

These balances data and their lineal regression can be plot in a graphic to see how linear the data is. The R^2 coefficient for b_1 and b_2 are 0.99 and 0.97, respectively.

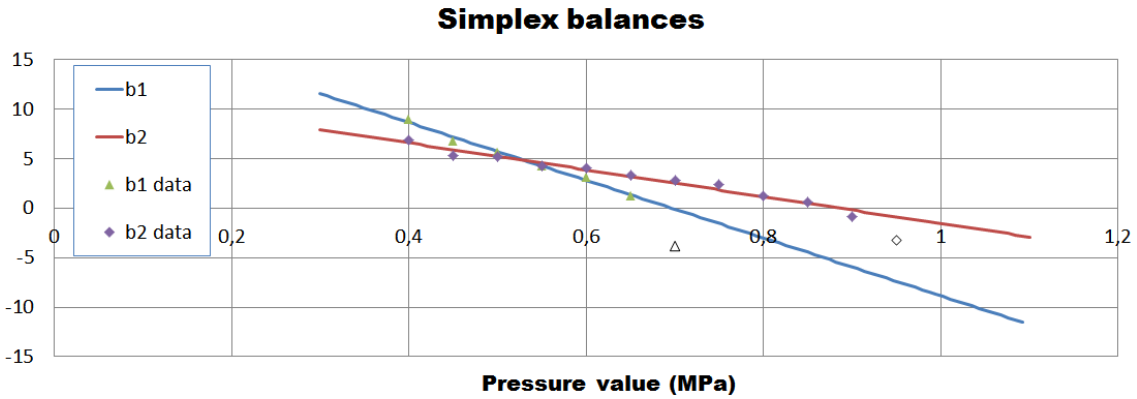


Figure 4. Linear regression of the logratio transformed data in the simplex space.

Finally, by retransforming this logratio values into the previous ones using ilr^{-1}, it is possible to obtain the probabilities of each possible output as a function of the overpressure.

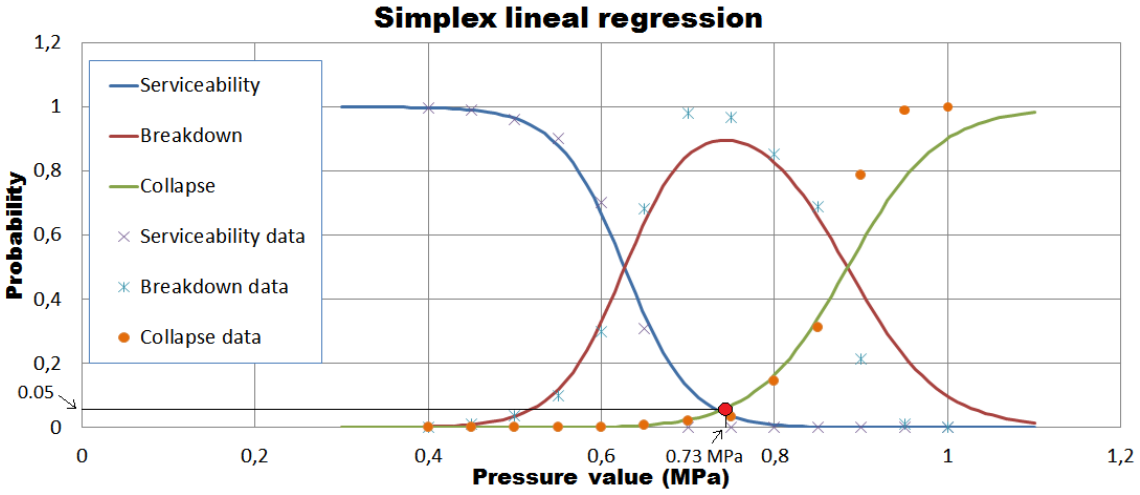


Figure 5. Final result, regression that shows the probability of each possible state after an overpressure event.

The failure criterion has been established as the pressure that causes collapse in the 5% of the cases. In this case, the inner pressure of failure is 0.7335 MPa. Other authors in Barbat et al. (1985)

have found that the overpressure value that follows this criterion is 1.05 MPa, which is slightly higher than the one obtained here. The reason is that even though the geometrical values and the constitutive properties that have been used are quite similar, in the present study, the prestressed tendons force have been considered in such a way that they simulate the force losses that occur gradually with time from 1985 to 2016 to update the results, so it is predictable that the supported overpressure here is lower. Moreover, in the present study the collapse scenario (Figure 3, in green), as stated previously, has been established as the one that produces a maximum principal tensile of 2 mm/m without taking into account that the reinforcing beams has a lot of resistance after reaching plasticity. On the other hand, in Barbat et al. (1985) the collapse criteria was fixed to be the one that cannot be resisted by the building itself. However, it has been here preferred to use the 2 mm/m strain value because this already implies huge cracks in the shell of the building that produce considerable leaks and so, failure.

Similar nuclear power plant containment buildings, for instance, Vandellós II, has been analysed in detail in the Stress Tests by Consejo de Seguridad Nuclear (CSN, 2011) and the pressure of failure has been determined to be 0.8667 MPa. From the results of the regression, it is possible to ascertain that this pressure would have a probability of serviceability of 0.0012, 0.5874 of breakdown, and 0.4114 of collapse.

Moreover, the pressure of design of Vandellós II (CSN, 2011) containment building was 0.3796 MPa. Dividing the pressure of failure obtained here (0.7335 MPa) by the pressure of design, the safety factor is 1.93, this is the main result of the analysis.

Finally, the large-break Loos of Coolant Accident (LOCA) has been considered traditionally as the worst scenario in which the containment building can be faced to. The effect of such a severe accident is typically taken as the Design Basis Accident (DBA). DBA is a postulated accident that a nuclear facility must be designed and built to withstand, it is necessary to assure public health and safety. A large-break LOCA can produce an inner pressure of 0.4 MPa, from the results of the regression, this pressure would have a probability of serviceability of 0.9972, 0.0028 of breakdown, and a really small one $2.7047 \cdot 10^{-7}$ of collapse, confirming the proper design.

Acknowledgments

Deepest gratitude to some Professors in the Civil Engineering School of the UPC who has helped and shown interest to this project, especially Dr. Antonio Aguado, Dr. Albert de la Fuente, and Dr. Antonio Rodriguez. They have offered invaluable support in all problems and their interest has been a source of motivation.

Many thanks as well to Associació Nuclear Ascó Vandellós for providing the data needed to develop this project and the Ministry of Education in Spain for providing the financial help through the Scholarship to collaborate with the Department of Applied Mathematics in UPC.

References

Aguado, A., A. Vives, J.J. Egozcue, and E. Mirambell (1991). Consideraciones sobre las bandas de tolerancia de la fuerza de pretensado en edificios de contención de centrales. 2as Jornadas Ibero-Latinoamericanas del Hormigón Pretensado. Buenos Aires, Argentina. pp. 481-508.

Aguado, A., J. M. Velasco, A. Vives, J.J. Egozcue, and E. Mirambell (1988). El ensayo de despegue y las bandas de tolerancia de la fuerza de pretensado en edificios de contención de centrales nucleares. Rev. Hormigón y Acero, nº 167. pp. 87-98.

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., C. Barceló-Vidal, J. 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. In Proceedings of IAMG'02 | The eigth annual conference of the International Association for

Mathematical Geology, Bayer, U., Burger, H., and Skala, W., editors, volume I and II, Selbstverlag der Alfred-Wegener-Stiftung, Berlin, 1106 p, 387-392.

Barbat A. H., M. Cervera, C. Cirauqui, A. Hanganu y E. Oñate (1995). Evaluación de la presión de fallo del edificio de contención de una central nuclear tipo PWR-W tres lazos. Parte 2: simulación numérica. Revista Internacional Métodos Numéricos para Cálculo y Diseño en Ingenierzá, Vol. 11, No. 3.

Cervera M., A. H. Barbat, A. Hanganu, E. Oñate y C. Cirauqui (1995). Evaluación de la presión de fallo del edificio de contención de una central nuclear tipo PWR-W Tres Lazos. Parte 1: Metodología. Revista Internacional de Métodos Numéricos para Cálculo y Diseño en Ingenieria, Barcelona, en prensa.

Consejo de Seguridad Nuclear (2011). Pruebas de Resistencia realizadas a las centrales nucleares españolas. Informe final tras el accidente de Fuckushima Dai-ichi.

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

Egozcue, J. J. and V. Pawlowsky-Glahn (2005). Groups of parts and their balances in compositional data analysis. Mathematical Geology, 37, 799-832.

Egozcue, J. J., J. Daunis, V. Pawlowsky-Glahn, K. Hron, P. Filzmoser (2012). Simplicial Regression. The normal model. Journal of Applied Probability and Statistics. Vol. 6, No. 1&2, pp. 87-108.

Hammersley, J. M. and D. C. Handscomb (1964). Monte Carlo Methods. Wiley, New York.

Jankowiak, T. and T. Lodyfowski (2005). Identification of parameters of concrete damage plasticity constitutive model. Publishing House of Poznan University of Technology, No. 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.

Hammersley, J. M. and Handscomb, D. C. (1964). Monte Carlo Methods. Wiley, New York.

Robert, C. and G. Casella (2004). Monte Carlo Statistical Methods. Springer, New York.

Spurious copulas

M.I. ORTEGO[1] and J.J. EGOZCUE[1]

[1] Departament de Matemàtica Aplicada III - Universitat Politècnica de Catalunya, Spain. ma.isabel.ortego@upc.edu

1 Introduction

Modeling dependence between two or more variables is a common issue in statistical applications. The Pearson correlation coefficient is often used to measure dependence, although it only captures linear dependence. Other dependence coefficients, such as Kendall's τ or Spearman's ρ, among others, model dependencies other than linear (Schweizer and Wolff, 1981; Scarsini, 1984).

1.1 Copulas

The use of this *other* coefficients of dependence is directly linked to the use of copula functions to model the multivariate variables at hand (Nelsen, 1999; Genest and Favre, 2007). The term copula in latin refers to connect or join. Copula functions join univariate cumulative distribution functions to obtain their multivariate CDF. Definitions and Theorems in this paper are referred to the bivariate case.

Definition A bivariate *copula* $C[u, v]$ is a function, $C : [0, 1]^2 \to [0, 1]$, which satisfies:

1. $C[u, 1] = u$, $C[1, v] = v$, for every u, v in $[0, 1]$.
2. $C[u, 0] = C[0, v] = 0$ for every u, v in $[0, 1]$.
3. $C[a_2, b_2] - C[a_1, b_2] - C[a_2, b_1] + C[a_1, b_1] \geq 0$, when $a_1 \leq a_2, \quad b_1 \leq b_2 \in [0, 1]$.

The use of copula functions allows to treat separately the marginal distributions of the variables and the dependence structure among them. Sklar's theorem (Sklar, 1959), ensures that under continuity hypothesis, the bivariate distribution between two variables can be expressed in a unique manner as a combination of the marginal distributions and their copula. Copulas seem then a useful tool to represent dependence between variables.

Theorem 1.1 (Sklar's Theorem) *If H is a bivariate CDF with univariate marginal CDFs F_1, F_2, then there exists a bivariate copula C that, for all x_1, x_2 in $\mathbb{R}$,*

$$H(x_1, x_2) = C[F_1(x_1), F_2(x_2)] \ .$$

If H is absolutely continous, then C is unique. Conversely, if C is a bivariate copula, and F_1, F_2 are univariate CDFs, then $H(x_1, x_2) = C[F_1(x_1), F_2(x_2)]$ is a bivariate CDF whose univariate marginals are F_1, F_2.

The most popular version of Th. 1.1, appears in Schweizer and Wolff (1981). This version of the theorem is expressed in terms of random variables:

Theorem 1.2 *If X_1, X_2 are real random variables, defined in a common probability space, with marginal CDFs F_{X_1}, F_{X_2} and joint CDF $H_{X_1X_2}$, then there exists a bivariate copula $C_{X_1X_2}$ so as to, for every x_1, x_2 in $\mathbb{R}$,*

$$H_{X_1X_2}(x_1, x_2) = C_{X_1X_2}[F_{X_1}(x_1), F_{X_2}(x_2)] \ .$$

If F_{X_1}, F_{X_2} are absolutely continous, then $C_{X_1X_2}$ is unique.

One of the most useful properties of copulas is their invariance for strictly monotonous transformations of the random variables:

Proposition 1.3 *(Invariance) Let X_1, X_2 be real random variables, with absolutely continous marginal CDFs F_{X_1}, F_{X_2} and copula C. Let T_1, T_2 be strictly increasing transformations. Then, the set of random variables $T_1(X_1)$,$T_2(X_2)$ has the same copula C that X_1, X_2.*

Therefore, the copula describes the form in which X_1, X_2 are related, independently of the scale in which each random variable is measured. Copulas capture nonparametric aspects of the relationship between variables, and therefore, association measures and dependence concepts are properties of the copula.Copula functions are usually represented through their density function.

1.2 Modeling dependence of compositional data

In a compositional framework, dealing with dependence is a key problem. In fact, the detection of spurious Pearson correlation can be considered as the spark that boosted the interest for compositional data and its suitable treatment. Other association coefficients have been defined. Spearman's ρ (Spearman, 1904; Kruskal, 1958) and Kendall's τ (Kendall, 1938) are coefficients based on concordance between variables:

Definition Given three independent and identically distributed random vectors, (X_1, X_2), (X_1^*, X_2^*), and (X_1^{**}, X_2^{**}) with joint CDF H and copula C. The *Spearman's rank correlation coefficient* is proportional to the concordance probability minus the discordance probability for the vectors (X_1, X_2) and (X_1^*, X_2^{**}):

$$\rho_S = 3(\mathrm{P}[(X_1 - X_1^*)(X_2 - X_2^{**}) > 0] - \mathrm{P}[(X_1 - X_1^*)(X_2 - X_2^{**}) < 0]) \ .$$

Kendall's rank correlation coefficient is also defined in terms of concordance:

Definition Given two independent and identically distributed random vectors, (X_1, X_2), (X_1^*, X_2^*), the *Kendall's rank correlation coefficient* is:

$$\tau(X_1, X_2) = \mathrm{P}[(X_1 - X_1^*)(X_2 - X_2^*) > 0] - \mathrm{P}[(X_1 - X_1^*)(X_2 - X_2^*) < 0] \ ,$$

a measure of the probability of concordace minus the probability of the discordance.

Concordance measures in general, and particularly Kendall's τ and Spearman's ρ, can be written in terms of the corresponding copula function (Schweizer and Wolff, 1981; Dupuis, 2007; Genest and Favre, 2007; Nelsen, 1999). On the other hand, these measures are invariant for strictly monotonous transformations of the variables, i.e, if κ_{X_1,X_2} is a measure of concordance and α and β are strictly monotonous functions (a.s.) over $\mathrm{Sup}(X_1)$ and $\mathrm{Sup}(X_2)$, respectively, then $\kappa_{\alpha(X_1)\beta(X_2)} = \kappa_{X_1,X_2}$.

Due to this invariance for strictly monotonous transformations, some researchers may expect that Kendall's τ, Spearman's ρ or the copula between parts of a composition are not spurious thus suggesting a way to circumvent log-ratio treatment of compositional data. In next section, these dependence coefficients and copula functions are used to measure the dependence between parts of a composition and between the parts of its subcompositions. It is shown that if the compositional structure of data is ignored, the use of these measures of dependence also presents great shortcomings, as they are subcompositionally incoherent and, consequently, they are spurious measures of dependence between parts of a composition.

2 Example

A set of compositional data (normal in the simplex: Aitchison and Shen (1980); Egozcue et al. (2012); Mateu-Figueras and Pawlowsky-Glahn (2008)) has been simulated (5 parts: $X[1]$ to $X[5]$ and 100 points). Fig. (1) shows the joint scatterplot of the parts. There are different degrees of dependence between parts, both positive and negative.

Three different dependence coefficients (Pearson, Spearman's ρ and Kendall's τ) have been computed to take account of the dependence between parts. Tables (1, 2) show the values of these coefficients for the sample. As in the scatterplot, different degrees of dependence are presented. We

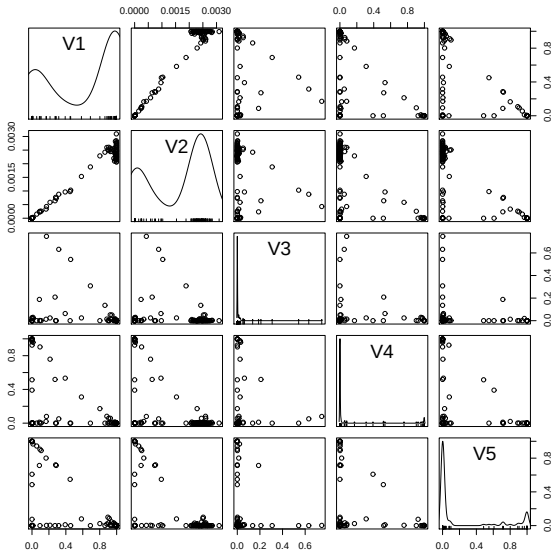


Figure 1: Scatterplot of the parts of the composition

center our attention on the dependence between parts $X[2]$ and $X[3]$ (red) and between parts $X[3]$ and $X[4]$ (blue). The value of the Pearson correlation coefficient for $X[2]$ and $X[3]$ suggests weak linear dependence between them, a value near to zero. When using Spearman's or Kendall coefficients, Table (2) values suggest medium dependence for the pair $X[3]$ and $X[4]$, and a very low dependence for $X[2]$ and $X[3]$.

	X[1]	X[2]	X[3]	X[4]	X[5]
X[1]	1.00	0.99	-0.15	-0.55	-0.67
X[2]	0.99	1.00	-0.14	-0.55	-0.66
X[3]	-0.15	-0.14	1.00	-0.03	-0.11
X[4]	-0.55	-0.55	-0.03	1.00	-0.19
X[5]	-0.67	-0.66	-0.11	-0.19	1.00

Table 1: Pearson correlation of the parts

	X[1]	X[2]	X[3]	X[4]	X[5]
X[1]	1.00	0.74	-0.20	-0.69	-0.70
X[2]	0.74	1.00	0.03	-0.40	-0.45
X[3]	-0.20	0.03	1.00	0.59	-0.06
X[4]	-0.69	-0.40	0.59	1.00	0.31
X[5]	-0.70	-0.45	-0.06	0.31	1.00

	X[1]	X[2]	X[3]	X[4]	X[5]
X[1]	1.00	0.58	-0.17	-0.51	-0.53
X[2]	0.58	1.00	0.02	-0.27	-0.31
X[3]	-0.17	0.02	1.00	0.41	-0.04
X[4]	-0.51	-0.27	0.41	1.00	0.19
X[5]	-0.53	-0.31	-0.04	0.19	1.00

Table 2: Spearman's and Kendall's correlation coefficients of the parts

In order to treat dependence between parts separately from the marginal distibutions, observations have been transformed into pseudo-observations based in ranks (Genest and Favre, 2007), a monotonous transformation that does not affect dependence. For parts $X[2]$ and $X[3]$ the corresponding scatterplot of pseudo-observations shows no particular dependence between parts (Fig. 2).

As Spearman's and Kendall's coefficients suggest a very low dependence between these parts, an independence test has been performed. The null hipothesis of independence cannot be rejected, and therefore parts $X[2]$ and $X[3]$ can be considered as independent.

As parts $X[3]$ and $X[4]$ present medium dependence, it seems adequate to represent this dependence through a copula function. First of all, an independence test has been performed, where the null

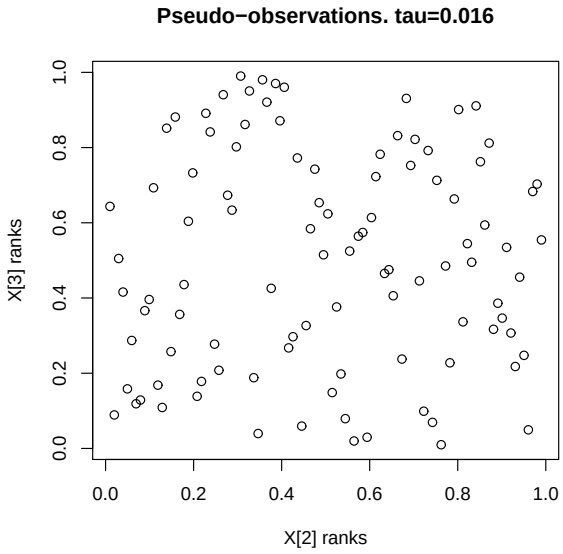


Figure 2: Scatterplot of pseudo-observations (rank-based) of X[2] and X[3]. Global Cramer-von Mises statistic: 0.034 with p-value 0.238

hypothesis has been rejected. Then, a set some well known family copulas have been assessed (see Table (5)), in order to choose the suitable ones for the pair. Fig. (3) shows the scatterplot of the corresponding pseudo-observations, as well as a set of suitable copula families for these data.

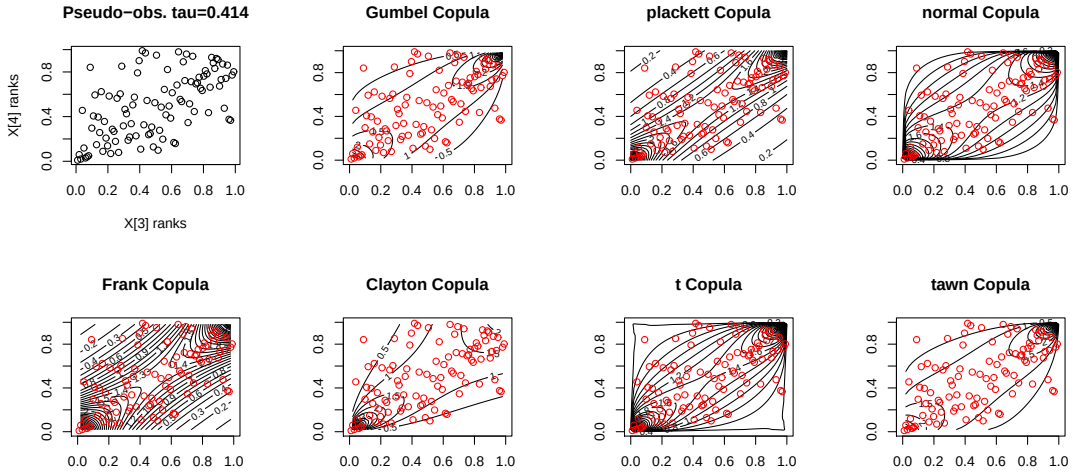


Figure 3: Scatterplot of pseudo-observations (rank-based) of X[3] and X[4] compared to the contours of the estimated copula density for some possible copula models

The tools used to treat compositional data must be subcompositionally coherent, i.e., they must show the same information if a subset of parts is used instead of the whole composition. Therefore, concordance coefficients and copulas are studied for subcompositions of the original data set. The subcomposition formed by parts $X[2]$ to $X[5]$ (i.e. supressing $X[1]$) is considered. Tables (3) and (4) show the values of the three coefficients (r,ρ,τ) for the parts of the compositions. Values for dependence between parts $X[2]$ and $X[3]$ are shown in red, and values for $X[3]$ and $X[4]$ are shown in blue.

At first view, the values of the Pearson correlation coefficient are spurious, as expected. The

	X[2]	X[3]	X[4]	X[5]
X[2]	1.00	-0.25	-0.43	-0.51
X[3]	-0.25	1.00	-0.12	-0.29
X[4]	-0.43	-0.12	1.00	-0.33
X[5]	-0.51	-0.29	-0.33	1.00

Table 3: Pearson correlation of the parts of subcomposition $X[2]$ to $X[5]$

	X[2]	X[3]	X[4]	X[5]
X[2]	1.00	0.16	-0.51	-0.49
X[3]	0.16	1.00	0.35	-0.37
X[4]	-0.51	0.35	1.00	-0.05
X[5]	-0.49	-0.37	-0.05	1.00

	X[2]	X[3]	X[4]	X[5]
X[2]	1.00	0.06	-0.37	-0.38
X[3]	0.06	1.00	0.22	-0.25
X[4]	-0.37	0.22	1.00	-0.06
X[5]	-0.38	-0.25	-0.06	1.00

Table 4: Spearman's and Kendall's correlation of the parts of subcomposition $X[2]$ to $X[5]$

changes in Spearman's and Kendall's coefficients are not as dramatic as those in Pearson's coefficient, but all values change. Therefore, even with good properties, as invariance for monotonous transformations, the concordance coefficients suffer from subcompositional incoherence.

In order to assess whether the independence of parts of $X[2]$ and $X[3]$ of the composition is mantained in the subcomposition, an independence test has been performed. The null hypothesis of independence is rejected, and therefore, parts $X[2]$ and $X[3]$ in the subcomposition can not be considered as independent. The scatterplot of the corresponding pseudo-observations show an slight dependence between both parts (Fig 4). Particularly, the profile of the upper right corner of the figure shows a clear dependence between the parts for the values in this region. This dependence can be modelled through a copula family. A set of some well known family copulas have been assessed. Fig. (5) shows the copula models of the set that have not been rejected in the assessment for this pair of variables. Parts $X[2]$ and $X[3]$ can be considered dependent or independent based on the manner in which they are measured. Therefore, copulas and tests are not working properly due to the compositional character of the data at hand.

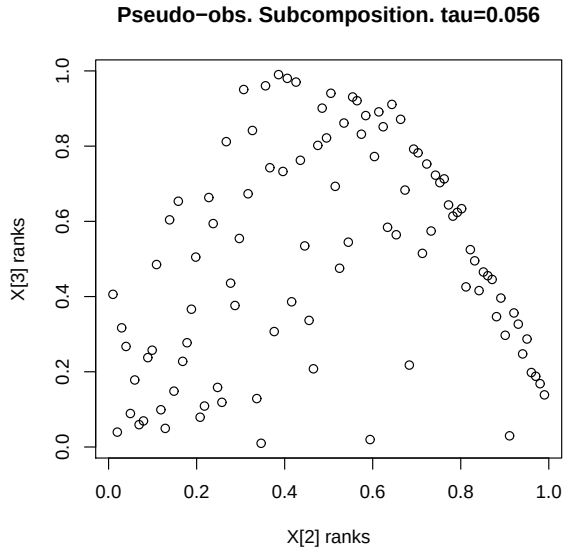


Figure 4: Scatterplot of pseudo-observations (rank-based) of X[2] and X[3] in the subcomposition. Global Cramer-von Mises statistic: 0.338 with p-value 0.0005

In order to assess if copula functions suffer also from subcompositional incoherence, the same set

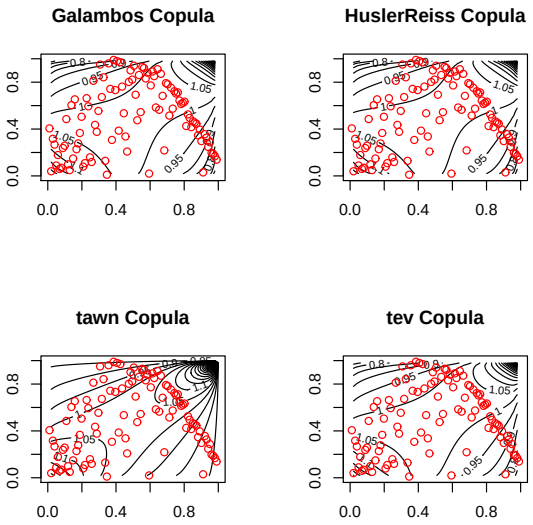


Figure 5: Scatterplot of pseudo-observations (rank-based) of X[2] and X[3] in the subcomposition compared to the contours of the estimated copula density for some possible copula models

of common copula families has been assessed to represent the dependence between parts X[3] and X[4] in the subcomposition. The subset of possible copula models for this dependence differs from the ones obtained to describe the dependence between parts X[3] and X[4] in the full composition (Table 5). Only two families seem suitable for both cases (Tawn and Clayton families). Even for these models, the differences between parameters suggest different strengths of the dependence of parts, when the only difference is the manner in which these parts have been measured. Therefore, although copulas seemed a promising tool, they also present compositional incoherence. Using copulas without considering the compositional structure of data could lead to errors.

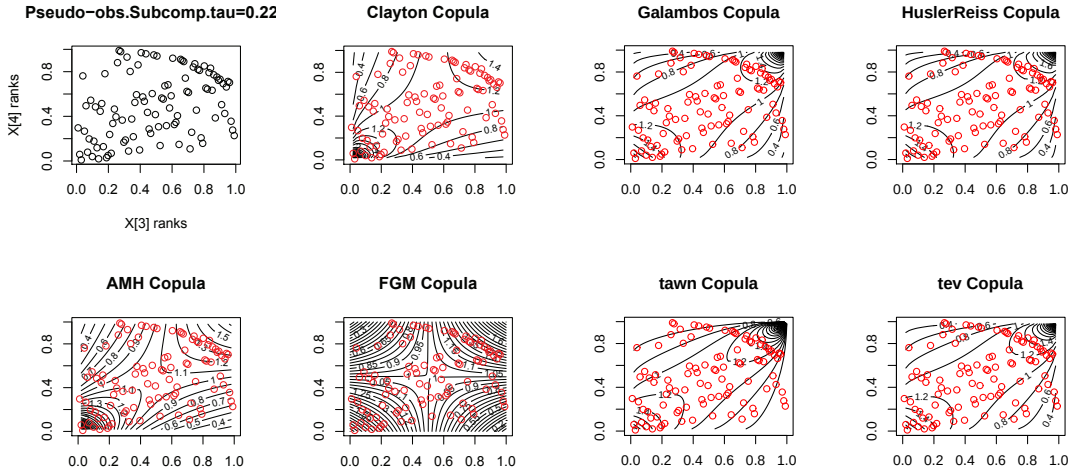


Figure 6: Scatterplot of pseudo-observations (rank-based) of X[3] and X[4] in the subcomposition compared to the contours of the estimated copula density for some possible copula models

Copula family \ Parameter	Composition	Subcomposition
Gumbel	1.707	–
Frank	4.359	–
Clayton	1.414	0.577
Plackett	7.089	–
Normal	0.606	–
t-copula	0.597	–
Tawn	0.992	0.591
AMH	–	0.776
FGM	–	1.0
Galambos	–	0.549
HuslerReiss	–	0.926
tev	–	0.558

Table 5: Set of assessed well known copula families. Values of the estimated parameters for parts $X[3]$ and $X[4]$. The sign – indicates that the copula family has been assessed for the pair and rejected.

3 Conclusions

The use of copulas or dependence coefficients other than the Pearson coefficient might seem a solution to treat compositional data. Spearman's ρ, Kendall's τ and several common copula models have been used to model the dependence between parts of a composition, and between the same parts in a subcomposition. The results show that copulas and concordance correlation coefficients suffer from subcompositional incoherence. Accordingly, these tools are only useful if the compositional character of data is taken into account. As a consequence, the dependence between parts of a composition expressed by copulas is spurious.

Acknowledgements

This research has received funding from the Spanish Government, projects COVARIANCE (CTM2010-19709), CODA-RSS (MTM2009-13272) and Metrics (MTM2012-33236).

References

Aitchison, J. and S. M. Shen (1980). Logistic-normal distributions. Some properties and uses. *Biometrika 67*(2), 261–272.

Dupuis, D. J. (2007). Using copulas in hydrology: Benefits, cautions and issues. *Journal of Hydrolgic Engineering 12*(4), 381–393.

Egozcue, J. J., J. Daunis-i-Estadella, V. Pawlowsky-Glahn, K. Hron, and P. Filzmoser (2012). Simplicial regression. the normal model. *Journal of Applied Probability and Statistics (JAPS) 6*(1–2), 87–108.

Genest, C. and A.-C. Favre (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *Journal of Hydrolgic Engineering 12*(4), 347–368.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika 30*, 81–93.

Kruskal, W. H. (1958). Ordinal measures of association. *Journal of the American Statistical Association 53*(284), 814–861.

Mateu-Figueras, G. and V. Pawlowsky-Glahn (2008). A critical approach to probability laws in geochemistry. *Mathematical Geosciences 40*(5), 489–502.

Nelsen, R. B. (1999). *An introduction to copulas*. New York, NY, USA: Springer-Verlag. 216p.

Scarsini, M. (1984). On measures of concordance. *Stochastica 8*(3), 201–218.

Schweizer, B. and E. F. Wolff (1981). On nonparametric measures of dependence for random variables. *The Annals of Statistics 9*(4), 870–885.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris 8*(1), 229–231.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology 15*(1), 72–101.

Nutrient balance ionomics: case study with mango (*Mangifera Indica*)

S.-É. PARENT[1], L.E. PARENT[1], D.E. ROZANE[2] and W. NATALE[3]

[1] ERSAM, Department of Soils and Agrifood Engineering, Université Laval, Québec, Canada, email: serge-etienne.parent.1@ ulaval.ca
[2]Departamento de Agronomia, Unesp, Universidade Estadual Paulista, Brazil
[3]Departamento de Solos e Adubos, Unesp, Universidade Estadual Paulista, Brazil

A plant ionome is the compositional vector of nutrients and trace elements in plant tissues. Ionomes and soil nutrients are commonly diagnosed in agronomy using concentration and nutrient ratio ranges. However, both diagnoses are biased by redundancy, scale dependency and non-normal distribution inherent to compositional data, potentially leading to conflicting results and wrong inferences. Our objective was to present an unbiased statistical approach of plant nutrient diagnosis using a balance concept and mango (Mangifera indica) as test crop. Ionomes were represented by balances computed as isometric log ratios (ilr). We collected foliar samples at flowering stage in 175 'Tommy', 'Palmer', 'Espada' and 'Haden' mango orchards in São Paulo state, Brazil. The ionomes comprised 11 nutrients (S, N, P, K, Ca, Mg, B, Cu, Zn, Mn, Fe). Soil fertility attributes (pH and bioavailable nutrients) were analyzed to reflect soil nutrient supply. Significant ($p < 0.05$) cultivar effect was confounded with soil effect in discriminant analysis and interpreted as high phenotypic plasticity of the species. A customized receiver operating characteristic (ROC) iterative procedure was developed to classify observation between true/false negative or positive and high/low-yielders. The ROC partitioning procedure showed that the critical Mahalanobis distance of 4.08 separating balanced from imbalanced specimens about yield cut-off of 128.5 kg fruit tree-1 proved to be a fairly informative test (area under curve = 0.84-0.92). Traditional multivariate methods were found to be numerically biased as shown by their deviation from the Mahalanobis distance of ilrs. We propose using a coherent pan balance diagnostic method with median ilr values of top yielders centered at fulcrums of a mobile and the critical Mahalanobis distance as a guide for global nutrient balance. Nutrient concentrations in weighing pans assist appreciating nutrients as relative shortage, adequacy or excess in balances.

Keywords: compositional data analysis, sequential binary partition, Cate-Nelson procedure, multivariate distance, mango cultivar, nutrient signature, ROC, mobile and fulcrums
Abbreviations: Acc.: accuracy; CND: Compositional Nutrient Diagnosis; DRIS: Diagnosis and Recommendation Integrated System; FN = false negative; FP = false positive; NPV = negative predictive value; PPV = positive predictive value; ROC = receiving operating characteristic; TN = true negative; TP = true positive.

1 Introduction

The ionome is "the mineral nutrient and trace element composition of an organism" (Lahner et al., 2003). Nutrient concentrations included in the definition of an ionome are strictly positive data constrained between zero and the unit of measurement: they thus belong to the compositional data class. Consequently, each nutrient can only be analyzed relatively to the other nutrients of the ionome. Ignoring the important properties of compositional data leads to numerical biases in their analysis due to redundancy, non-normal distribution and scale dependency (Bacon-Shone, 2011).

The ionome of agricultural crops is typically diagnosed using critical nutrient concentration ranges (CNCR) (Benton Jones et al., 1991) or dual ratios possibly integrated into functions and indices by the Diagnosis and Recommendation Integrated System (DRIS) (Walworth and Sumner, 1987). The CNCR and DRIS are usually conducted separately, and then compared to each other to identify the most limiting nutrients (Wadt and Silva, 2010). The CNCR, inherited from Sprengel's 'Law of minimum' stated in 1828, classifies crop nutrient as deficiency, sufficiency, luxury consumption, or excess (Epstein and Bloom, 2005). DRIS is an empirical model (without mathematical or statistical theory behind it) that computes $D\times(D\text{-}1)/2$ dual ratios and their associated ratio functions, then integrates functions into D indexes, although a composition has rank D-1 (Aitchison and Greenacre, 2002). DRIS was rectified by Parent and Dafir (1992) into a mathematically sound model using the centered log ratio (*clr*) transformation of Aitchison (1986). Although the *clr* has adequate geometry in the Euclidean space, it produces a singular matrix in multivariate analysis and one variate must be sacrificed to achieve D-1 degrees of freedom.

Obviously, the most publications in ionomics (e.g. Baxter et al., 2008, Gang et al. 2013) disregard the special properties of compositional data. An appropriate transformation is necesssary to avoid numerical

biases. Egozcue et al. (2003) proposed the isometric log-ratio transformation, which structures *D* components as *D*-1 orthogonal contrasts of components amenable to multivariate analyses. In plant nutrition, these contrasts can be defined as ad hoc nutrient balances. In a functional perspective, nutrient concentrations interact (Wilkinson et al., 2000) within in a structured system partitioned into subsystems (Marschner, 1995). Such balance concept has been used in a discriminant analysis computed across to represent the ionomes of plant species and varieties in a dimensionally reduced space (Parent et al., 2013). Also, nutrient imbalance indexes have been computed for diagnostic purposes as a distance between an observation and the center of a reference group of balanced specimens (Parent et al., 2012).

Our objectives were (1) to demonstrate numerical biases in traditional concentration and ratio methods to diagnose ionomes in agronomy, (2) to present a binary classification statistical technique developed to define a reference group and (3) to elaborate a pan balance representation of the ionome for nutrient diagnosis using mango as test plant.

2 Theory

2.1 Compositional data

Compositional data, such as nutrient concentrations, are part of some whole, or bounded between 0 and the unit of measurement, i.e. as 1, 100%, 1000 g kg^{-1}, or 10^6 mg kg^{-1}. The scale of measurement is generally the dry matter weight basis (kg), but could also be fresh matter weight basis (kg) or the sap liquid basis (L). The constrained nature of compositional data implies important numerical properties, as follows:

- *Redundancy*: The amount of one component can be calculated by difference between the measurement unit and the sum of the others. Hence, **there are *D*-1 degrees of freedom in a *D*-parts composition**, i.e. the data set matrix has rank *D*-1 (Aitchison and Greenacre, 2002). Because any of the *D*×(*D*-1)/2 dual ratios can be computed from other ratios (Parent et al., 2012), they convey redundant information that generates myriads of spurious correlations in linear statistical analysis (Pearson, 1897; Chayes, 1960).
- *Scale dependency*: The results of statistical tests differ depending on scale of measurement. **Scale depends on the way a composition is defined**, which generates spurious correlations (Tanner, 1949). Indeed, the addition of a component such as water to the composition just provides an additional dimension to that space (here water content). This new component should not change the results of the statistical analysis of a coherent, scale-invariant, system.
- *Non-normal distribution*: Normally distributed data are mapped in a real space, which is not the case for compositional data, which are mapped in a closed space. Statistics like confidence intervals should not be allowed to range outside the limit of the compositional space (e.g. ≤0% or ≥100%). Rather, **compositional data follow logistic-normal distributions** (Bacon-Shone, 2011).

Those three numerical properties are a consequence of closing the compositional space as follows (Aitchison, 1986):

$$S^D = C(c_1, c_2, \dots, c_D) = \left(\frac{c_1 \kappa}{\sum_{i=1}^{D} c_i}, \frac{c_2 \kappa}{\sum_{i=1}^{D} c_i}, \dots, \frac{c_D \kappa}{\sum_{i=1}^{D} c_i}\right) \quad \text{Eq. 1}$$

Where S^D is the simplex (compositional vector mapped in the compositional space), κ is the unit of measurement and c_i is the i^{th} part of a composition containing *D* parts. When conducting plant nutrient diagnosis, it is convenient to include a filling value (*Fv*) computed by difference between κ and the sum of all nutrients. Its inclusion allows to back-transform the *ilr* values into more familiar units of measurement. The main components of the filling value are C, O and H, as found in products of photosynthesis.

2.2 Conventional approaches

2.2.1 Nutrient concentration ranges

The critical nutrient concentration ranges (CNCR) has been illustrated by the so-called Liebig's barrel filled with water, where nutrient concentrations are represented by staves of unequal length, the shortest stave being attributed to the most limiting nutrient. The rise of water in the barrel, a metaphor for plant growth, is controlled by the shortest stave. Diagnosticians who use this approach interpret concentrations as nutrient deficiency, sufficiency, luxury consumption or excess. However, such approach does not account for nutrient interactions. Concentration data are often ordinary log transformed to improve data distribution. However, the ordinary log transformation is not a panacea (Filzmoser et al., 2009), because there are still *D* variables for matrix rand *D*-1, scale dependency is maintained and even though data are constrained into the positive space, they are not bounded by the unit of measurement.

2.2.2 Nutrient ratio ranges

Nutrient ratios have a long tradition in agronomy, aiming to capture the notion of nutrient interactions (Walworth and Sumner, 1987). Useful ratios are generally examined by data-mining, looking for high correlations with a performance index such as crop yield. However, since Pearson (1897), many statisticians (Tanner, 1949; Chayes, 1960, Aitchison, 1986) warned that the use of unstructured, correlation driven, dual ratio generates spurious correlations.

2.2.3 Nutrient stoichiometric rules

Ingestad (1987) suggested an optimum N:P:K:Ca:Mg stoichiometric rule for regulating the growth of tree seedlings, leading to ratios of all nutrients against a standard, e.g., N. This approach structures ratios in a way to avoid redundancy and scale-dependency, but still could lead to wrong interpretations namely due to non-normal distributions. The compositional formulation of stoichiometric rules is the additive log ratio transformation of Aitchison (1986) (see below).

2.2.4 Diagnosis and Recommendation Integrated System (DRIS)

The DRIS is a method to synthesize several dual ratios. Dual ratios in a given ionome are first compared to DRIS dual ratio norms (ratios obtained from high-performing specimens) to compute DRIS functions. The DRIS functions common to a nutrient are then added up to DRIS indices with the sign of DRIS functions depending on the position of the indexed nutrient in the ratio. Details of DRIS computations can be found in Walworth and Sumner (1987). Although appealing to plant diagnosticians, DRIS has poor mathematical ground. Parent and Dafir (1992) rectified DRIS for plant diagnosis using the *clr* transformation of Aitchison (1986).

2.3 Compositional approaches

2.3.1 The additive and centered log ratios (*alr*)

The *alr* representation of compositional data is computed as follows (Aitchison, 1986):

$$alr_i = ln\left(\frac{c_i}{c_{com}}\right) \qquad \text{Eq. 2}$$

Where c_i is the i^{th} component at numerator, $i = [1 \ldots D] \setminus i_{com}$ and c_{com} is the common denominator to all components, resulting in D-1 *alr* values, because the component at denominator is sacrificed. Log-ratios are more tractable than ordinary ratio (Aitchison, 2005), because the inverse of a log-ratio is a trivial sign change. The *alrs* are appropriate to conduct multivariate analysis, but are not linearly independent from each other, making them difficult to interpret. Distance-based statistics across additive dual ratios are not recommended (van den Boogaart et al., 2013).

2.3.2 The centered log ratios (*clr*)

The *clr* representation of compositional data is computed as follows (Aitchison, 1986):

$$clr_i = ln\left(\frac{c_i}{g(c)}\right) \qquad \text{Eq. 3}$$

Where c_i is the i^{th} component at numerator, $i = [1 \ldots D]$, and $g(c)$ is the geometric mean of all components, resulting in D *clr* values, i.e. there is one extra degree of freedom in a matrix of rank D-1. The *clr* transformation generates a singular matrix (the *clr* variates sum up to 0), one *clr* value must be sacrificed (e.g. that of the filling value) in multivariate analysis. Because outliers may affect considerably log ratios (Filzmoser and Gschwandtner, 2013), the diagnostic power of CND-*clr* is decreased by large variations in nutrient levels (e.g. leaf Cu, Zn, Mn contamination by fungicides). Nevertheless, the *clr* transformation is useful to conduct exploratory analyses on compositional data (Egoscue and Pawlowsky-Glahn, 2011).

2.3.3 The isometric log ratio (*ilr*)

The *ilr* technique (Egozcue et al., 2003) allows projecting the simplex S^D of compositional data into a Euclidean space of D-1 non-overlapping orthogonal log-contrasts, also called orthonormal balances or "coordinates". A system of balances can be designed in a sequential binary partition (SBP). A SBP is a (D-1)×D matrix, in which parts labeled "+1" (group numerator) are contrasted with parts labeled "-1" (group denominator) in each ordered row (see Table 1 for an example). A part labeled "0" is excluded from the balance. The composition is partitioned sequentially at every ordered row into two contrasts until the (+1) and (-1) subsets each contain a single part. In this paper, balances are conventionally noted as [-1 group | +1 group]. Balances are computed as follows (Egozcue and Pawlowsky-Glahn, 2005):

$$ilr_j = \sqrt{\frac{n_j^+ n_j^-}{n_j^+ + n_j^-}} ln \frac{g(c_j^+)}{g(c_j^-)} \quad \text{Eq. 4}$$

Where, in the j^{th} row of the SBP, n_j^+ and n_j^- are the numbers of components in the plus (+) or group and the minus (-) or group, respectively, $g(c_j^+)$ is the geometric mean of components in the + group and $g(c_j^-)$ is the geometric mean of components in the - group. The natural log of the ratio of geometric means is a log-contrast; the associated coefficient is an orthogonal coefficient assuring orthogonality or linear independence to *ilr* coordinates. The *ilr* transformation is appropriate for robust multivariate analysis of compositional data (Filzmoser and Hron, 2011).

2.3.4 Designing a sequential binary partition (SBP)

Wilkinson et al. (2000) listed several dual and higher-order nutrient interactions in plants in much larger numbers than the *D*-1 linearly independent variables allowable from a *D*-part composition (Aitchison and Greenacre, 2002). We thus designed a sound SBP for plant ionomes (Table 1). Nutrients were first contrasted with the filling value computed by difference between unit of measurement and the sum on nutrient concentrations. Macro-nutrients and B were separated from cationic micronutrients. Macro-nutrients and B are connected, because B interacts with macronutrients (Malavolta, 2006). Macro-nutrient anions (S, N, P) were contrasted with macro-nutrient cations (K, Ca, Mg) to reflect charge balance in plant cells. Macro-nutrient anions were further subdivided according to protein synthesis (N, S) and energy (P); the [P | S,N] balance reflects the protein/energy relationship in plants. Macro-nutrient cations were contrasted as monovalent vs. divalent ions whereby K, Ca and Mg are competing nutrients (Marschner, 1995). Fungicides that protect agricultural plants against diseases often contain Cu, Zn and Mn in their active molecules. The Cu and Zn were thus assumed to be mainly affected by fungicide sprays; Mn may originate from soil or fungicide sprays while soil can be assumed to be the large reservoir of Mn and Fe; the [Cu,Zn,Mn | Fe] balance is intended to reflect the effect of fungicide sprays over soil supply of cationic micronutrients.

ILR ID	S	N	P	K	Ca	Mg	B	Cu	Zn	Mn	Fe	Fv	Notation
1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	+1	-1	[Fv | S,N,P,K,Ca,Mg,B,Cu,Zn,Mn,Fe]
2	+1	+1	+1	+1	+1	+1	+1	-1	-1	-1	-1	0	[Cu,Zn,Mn,Fe | S,N,P,K,Ca,Mg,B]
3	+1	+1	+1	+1	+1	+1	-1	0	0	0	0	0	[B | S,N,P,K,Ca,Mg]
4	+1	+1	+1	-1	-1	-1	0	0	0	0	0	0	[K,Ca,Mg | S,N,P]
5	+1	+1	-1	0	0	0	0	0	0	0	0	0	[P | S,N]
6	+1	-1	0	0	0	0	0	0	0	0	0	0	[N | S]
7	0	0	0	+1	-1	-1	0	0	0	0	0	0	[Ca,Mg | K]
8	0	0	0	0	+1	-1	0	0	0	0	0	0	[Mg | Ca]
9	0	0	0	0	0	0	0	+1	+1	-1	-1	0	[Fe | Cu,Zn,Mn]
10	0	0	0	0	0	0	0	+1	-1	0	0	0	[Mn | Cu,Zn]
11	0	0	0	0	0	0	0	0	0	+1	-1	0	[Zn | Cu]

Table 1. Sequential orthogonal partition of eleven nutrients of plant ionome and the filling value to compute 11 *ilr* orthonormal coordinates from concentration values and orthogonal coefficients

2.3.5 Dissimilarity between two compositions

The Mahalanobis distance ($\mathcal{M}$) across selected *ilr* coordinates of ionomes is computed as follows:

$$\mathcal{M} = \sqrt{(x-\bar{x})^T COV^{-1}(x-\bar{x})} \quad \text{Eq. 5}$$

Where $\bar{x}$ is the barycentre of a reference population and COV is the covariance matrix of the reference population. The Mahalanobis distance across *ilrs* is a measure of the multivariate distance between a diagnosed and a reference composition. The Mahalanobis distance can thus be used as multivariate index of nutrient imbalance.

2.4 Binary classification method

For diagnostic purposes, there is a need to split the crops into low- and high-productivity groups. A predictor index should allow separating balanced from unbalanced nutrient signatures. Four quadrants are partitioned in system diagnosis (Swets, 1988), where each quadrant defines a response class (Table 2) according to response and predictor delimiters. The Mahalanobis distance is computed from the centre and the covariance of the reference population. To define an optimal predictor delimiter, Nelson and Anderson (1977) proposed to maximize the "Class sum of squares" between two groups clustered by the predictor delimiter for a given response delimiter.

In the Receiver Operating Characteristic (ROC) method, the selected predictor delimiter corresponds to the best compromise between sensitivity and specificity, i.e. the maximal value of sensitivity × specificity. In

other words, the optimal delimiter is the one corresponding to the nearest point to the (1,1) corner of the sensitivity versus specificity plot. The area under the sensitivity versus specificity curve (AUC) can also be used as an accuracy index for the partition (Swets, 1988). Because crop yield is a continuous variable, a procedure is needed to optimize the response delimiter, as developed in the Material and methods section.

In survey datasets, true negative (TN) specimens represent the reference population. Because the Mahalanobis distance is used as predictor for TNs ($\mathcal{M}_{TN}$), an iteration procedure is needed. For a given response (crop yield) delimiter, the predictor is initiated using high-yielders as reference specimens for computing $\mathcal{M}_{HY}$. Thereafter, a predictor delimiter is selected and its barycenter and covariance are computed among newly delineated TN specimens to solve $\mathcal{M}_{TN}$. The $\mathcal{M}_{TN}$ is iterated until two iterations classifies observations identically. The Moore-Penrose pseudo-inversion was used to avoid singularities in the inversion of the covariance matrix (Prekopcsák and Lemire, 2012).

Response		
Negative predictive value (NPV): probability that a balance diagnosis returns high yield, as TN/(TN+FN)	**Positive predictive value** (PPV): probability that an imbalance diagnosis returns low yield, as TP/(TP+FP)	**Accuracy**: probability that an observation is correctly diagnosed as balanced or imbalanced, as (TP+TN)/(TP+TN+FP+FN)
True negative (TN: nutrient balance): high yield crops correctly diagnosed as balanced (below predictor critical index). The nutrient status of the plant is adequate.	**False positive** (FP: type I error): high yield crops incorrectly identified as imbalanced (above predictor critical index). FP observations indicate luxury consumption of nutrients by the plant.	**Specificity**: probability that a high yield is balanced, as TN/(TN+FP)
False negative (FN: type II error): low yield crops incorrectly identified as balanced (below critical index). FN observations indicate the impact of other limiting factors on crop performance.	**True positive** (TP: nutrient imbalance): low yield crops correctly diagnosed as imbalanced (above critical index). At least one nutrient is imbalanced.	**Sensitivity**: probability that a low yield is imbalanced, as TP/(TP+FN)

Response delimiter

Predictor delimiter — Predictor

Table 2. Term définitions and their relationships in plant balance binary classifications

3 Material and methods

3.1 Data set

We collected data in 93 'Palmer'. 63 'Tommy', 14 'Espada', and 5 'Haden' mango orchards planted between 1983 and 2005 on Oxisols and Ultisols near Jabotocabal in the state of São Paulo, Brazil. At the end of July during flowering, leaves were collected from the middle tier of annual growth. Foliar N was determined by micro-Kjeldahl. The S, P, K, Ca, Mg, Zn, Cu, Mn, Fe, and B foliar concentrations were determined by IPC-OES after digestion in a mixture of nitric and perchloric acids (Jones and Case, 1990). Fruits were harvested from five trees randomly selected in each orchard, and averaged as kg tree^{-1}.

Phenotypic plasticity of plant ionome may be driven by nutrient supply of soils. Soils were thus sampled after harvest at four locations per tree in the 0-20 cm layers, then composited per 5-tree experimental unit. Soil samples were air dried and analyzed for pH in 0.01 M $CaCl_2$, organic matter content, P, K, Ca, Mg and (H + Al), and micro-nutrients using Brazilian methods (Raij et al., 1987). Exchangeable acidity (H+Al) was determined by the SMP pH buffer method and the equation of Quaggio et al. (1985) to convert buffer pH to $mmol_c$ (H+Al) dm^{-3} as follows:

$(H + Al) = 10\exp(7.76 + 1.053pH_{SMP})$, $R^2 = 0.98$ Eq.6

3.2 Classification of nutrient balances

The Mahalanobis distance from the median of the TN specimens was used as predictor. The response (productivity criteria) delimiter returning the largest area under ROC curve (AUC) was selected. For

statistical validity, the delimiters associated to maximum AUC should also include sufficient data classified in the reference population (TN). Because the amount of data is relatively small, we retained the minimum of 20 observations in the TN quadrant.

3.3 Statistical analysis

Statistical computations were conducted in the R statistical environment (R Development Core Team 2013) using the R "compositions" package (van den Boogaart et al., 2013). Outliers among *ilrs* were discarded at the 0.01 level using the R "mvoutlier" package (Filzmoser and Gschwandtner, 2013). We compared *ilr* coordinates of ionomes using Tukey's test at a 0.05 significance level. Variances of balances were compared using Bartlett's test and their mean were compared using analysis of variance ($p \leq 0.05$). Because tests are multivariate and plant data sets contain extreme values, a robust method based on the median is needed to compute multivariate distances (Filzmoser et al., 2009).

4 Results

4.1 Cultivar ionomes

Bartlett test showed that the variance of 3 of the 11 balances differed among varieties, i.e. [Fv | nutrients], [B | S,N,P,K,Ca,Mg] and [N | S]. Analysis of variance showed that 8 of the 11 balance means differed among varieties, with the exception of [B | S,N,P,K,Ca,Mg] (barely interpretable due to heterogeneous variance), [Mn | Cu,Zn] and [Zn | Cu]. The discriminant scores mapped the differences between ionomes of 'Palmer', 'Tommy', 'Espada' and 'Haden' (Figure 1). The plant and soil DA maps showed that ionomes differed between varieties. We could not reject the hypothesis of phenotypic plasticity because the effects of soil conditions and varieties were confounded. Because the genotype effect could not be ascertained, the ROC and Cate-Nelson partitioning procedures were run across varieties.

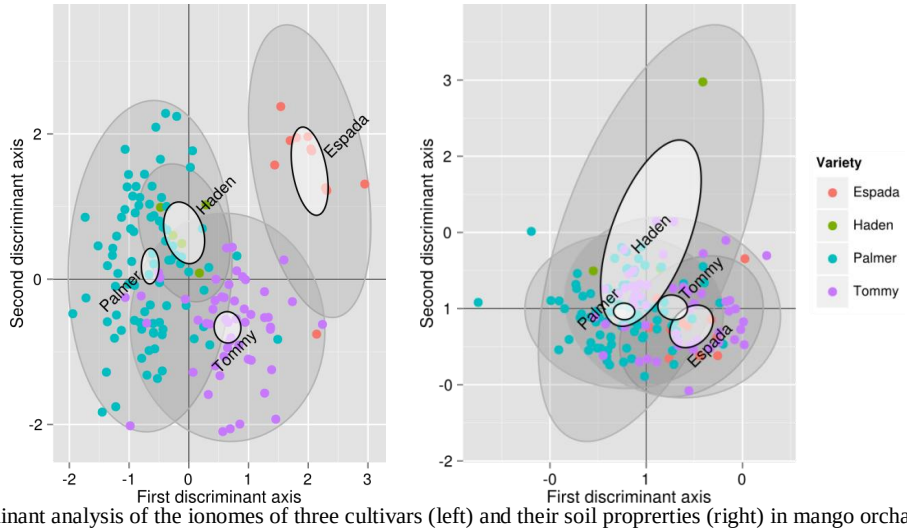
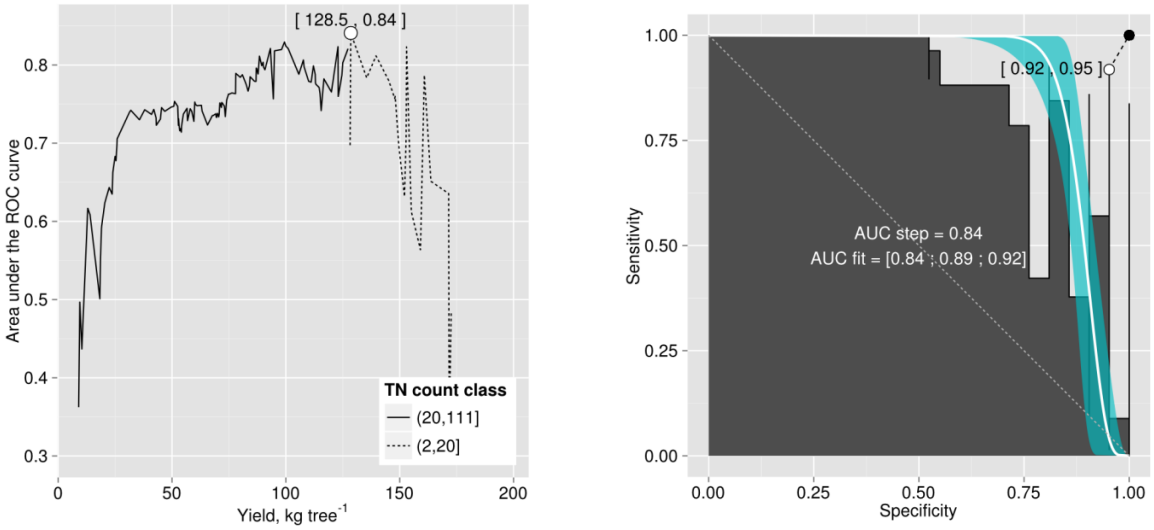


Figure 1. Discriminant analysis of the ionomes of three cultivars (left) and their soil proprerties (right) in mango orchards in the state of São Paulo, Brazil: 'Palmer' (93 obs.), 'Tommy' (63 obs.), 'Espada' (14 obs.) and Haden (5 obs.). Large semitransparent ellipses that enclose swarms of data points represent regions that include 95% of the theoretical distribution of canonical scores for each species. Smaller plain white ellipses represent confidence regions about means of canonical scores at 95% confidence level.

4.2 Binary classification

The area under the ROC curve (AUC), computed by summing rectangles under the step curve, reached a peak at 0.84 (Figure 2a), a value comparable to the AUC for fairly informative tests (0.80-0.98) in medical sciences (Swets, 1988). The ROC curve did not show a regular decrease of sensitivity as specificity increased, as usually observed in ROC diagnoses due to the re-sampling of the TN specimens (see methodology section), which is generally not needed in conventional clinical studies. The AUC computed across the fitted binormal model (Hadley, 1988) returned a value of 0.89, ranging between 0.84 and 0.92, with a confidence level of 0.95 (Figure 2b). The response delimiter corresponding to the AUC peak was 128.5 kg tree^{-1} (Figure 2a). The optimal compromise between specificity and sensitivity for the optimal response delimiter was found at a specificity of 0.95 and a sensitivity of 0.92, corresponding to a predictor threshold (Mahalanobis distance) of 4.08 (Figure 2b).



(a) (b)

Figure 2. (a) area under the ROC curve versus cut-off yield and (b) ROC curve for yield cut-off of 128.5 kg fruit tree^{-1}.

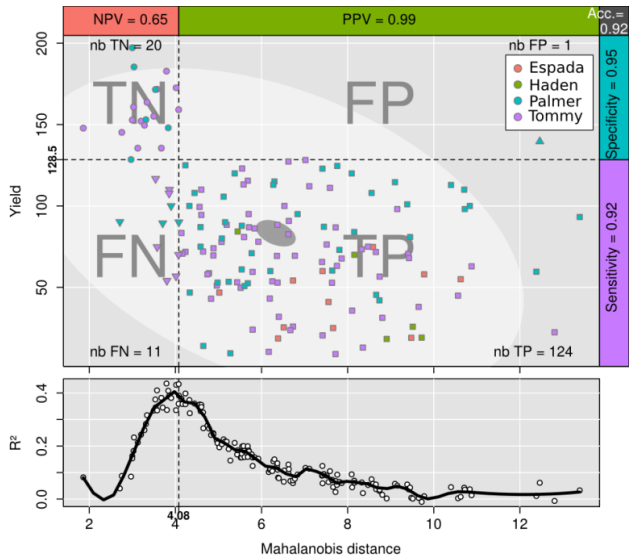


Figure 3. Binary classification of data with indexes (top). (bottom) Class sum of squares (bottom).

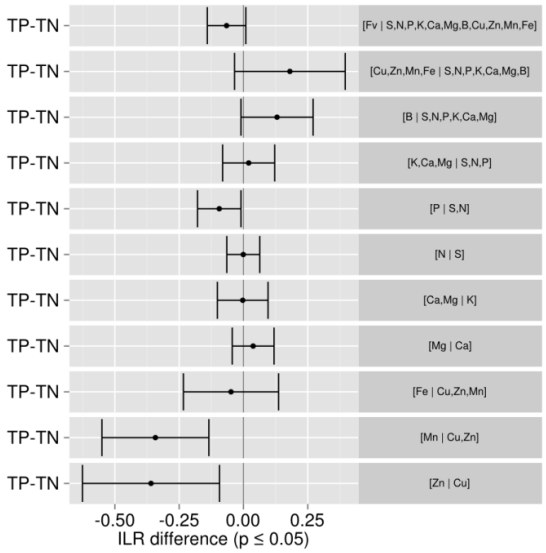


Figure 4. Tukey test of *ilr* differences between TP and TN specimens.

Results of the binary classification are presented in Figure 3, where the two optimal delimiters set apart 20 observations in the TN quadrants. The semi-transparent ellipse enclosed 95% of the theoretical distribution of all observations. The TN group was essentially constituted of 'Tommy' and 'Palmer' orchards. All 'Espada' and 'Haden' orchards were classified as TP. The large majority (accuracy = 92%) of specimens were correctly diagnosed by the Mahalanobis distance predictor. Almost all specimens declared imbalanced yielded less than cut-off yield value (PPV = 99%). On the other hand, nearly two thirds (NPV=65%) of balanced specimens yielded more than cut-off yield value.

Median *ilr* values of TN specimens as well as the covariance matrix used to measure Mahalanobis distances are presented in Tables 4 and 5.

4.3 Nutrient balance comparisons between TN and TP specimens

Tukey's test allowed detecting in which balance significant differences occurred between TN and TP specimens (Figure 4). The most significantly different balance contrast was [Mn | Cu,Zn] ($p < 0.01$). A negative (TP-TN) value means that TN's balance is higher than TP's. On the [Mn | Cu,Zn] balance, compared to TP, the TN specimens tend to be characterized by greater load in the + group compared to the - group. In this case, Cu and Zn loaded more than Mn in the TN ionomes. There was a significant trend for TN specimens to accumulate more Cu relatively to Zn, as shown by the negative [Zn | Cu] balance difference

(TP-TN) ($p < 0.01$). Finally, compared to TP specimens, TN specimens accumulated significantly more S and N than P (i.e TN had a higher [P | N,S] balance) ($p < 0.05$). There was no significant difference between other balances.

	LL	Median	UL
[Fv | S,N,P,K,Ca,Mg,B,Cu,Zn,Mn,Fe]	-7.153	-7.086	-7.000
[Cu,Zn,Mn,Fe | S,N,P,K,Ca,Mg,B]	5.394	5.518	5.859
[B | S,N,P,K,Ca,Mg]	4.454	4.639	4.720
[K,Ca,Mg | S,N,P]	-1.312	-1.205	-1.170
[P | S,N]	1.227	1.300	1.338
[N | S]	-1.652	-1.628	-1.551
[Ca,Mg | K]	0.248	0.345	0.407
[Mg | Ca]	1.563	1.598	1.682
[Fe | Cu,Zn,Mn]	-0.162	-0.005	0.095
[Mn | Cu,Zn]	-2.554	-2.374	-2.131
[Zn | Cu]	-0.148	0.218	0.539

Table 4. Confidence intervals of *ilr* values ($\pm t_{0.025}\sqrt{s^2/n}$) for true negative (TN) specimens (n = 20) in the Brazilian mango data set (LL = lower limit; UL = upper limit)

	ilr1	ilr2	ilr3	ilr4	ilr5	ilr6	ilr7	ilr8	ilr9	ilr10	ilr11
ilr1	0.0267										
ilr2	-0.0653	0.2467									
ilr3	-0.0289	0.0403	0.0808								
ilr4	0.0158	-0.0299	-0.0173	0.0233							
ilr5	0.0068	-0.0344	0.0024	-0.0008	0.0140						
ilr6	0.0076	-0.0233	-0.0025	0.0009	-0.0003	0.0117					
ilr7	0.0053	0.0076	0.0000	0.0109	-0.0092	0.0038	0.0288				
ilr8	-0.0032	0.0003	0.0132	-0.0082	0.0024	0.0039	-0.0039	0.0161			
ilr9	0.0203	-0.0612	-0.0388	0.0196	0.0056	0.0024	0.0003	-0.0106	0.0757		
ilr10	0.0507	-0.1155	-0.0809	0.0330	-0.0048	0.0266	0.0219	-0.0091	0.0617	0.2040	
ilr11	0.0619	-0.2338	-0.0614	0.0111	0.0423	0.0156	-0.0330	-0.0254	0.0626	0.1599	0.5378

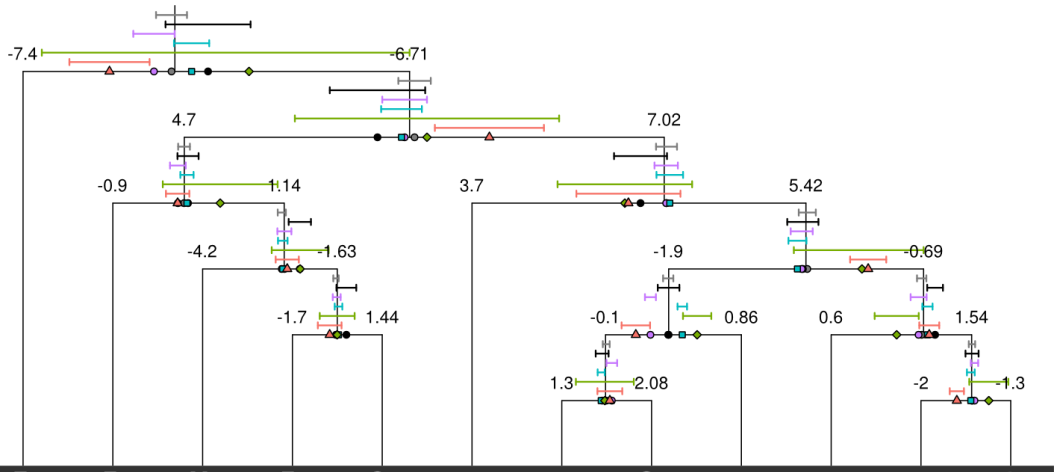
Table 5. Covariance matrix (excluding outliers) of TN specimens to compute the Mahalanobis distance for mango observations in the Brazilian data set

4.4 Pan balance diagram

Balances can be represented metaphorically using a stand-alone mobile diagram with fulcrums and weighing pans, where nutrient concentrations in buckets impact directly on nutrient balances at fulcrums upon change. Figure 5 presents a balance dendrogram derived from SBP with overall average *ilr* values at fulcrums, and 0.05 confidence intervals for TN specimens, TP specimens, and each cultivar.

The *ilr* values at fulcrums are used for diagnostic purposes, while the *ilr* values back-transformed to familiar concentration units are laid down in weighing pans to provide an appreciation of balances in terms of relative shortage, adequacy, luxury consumption or excess of contributing nutrients. Differences between TN and TP specimens can be observed in the pan balance diagram. There were marked differences between the TNs and TPs in Cu and concentrations, apparently misbalancing significantly [Mn | Zn,Cu] and [Zn | Cu]. Although P shortage seemed to be small in TP specimens, it contributed to misbalance [P | N,S] in TP specimens.

Large confidence intervals for 'Haden' were due to too few observations (5). The departure from the TN [Fe | Mn,Zn,Cu] fulcrum, although not significant, is attributable to a balance driven positively by relatively low Fe and high Cu levels. The 'Espada' [Fe | Mn,Zn,Cu] balance was similar to that of 'Tommy' and 'Palmer'. Even though 'Espada' showed relatively low Fe, Mn, Zn and Cu levels, this apparent shortage of nutrients was properly balanced. However, those low concentrations were misbalanced with other nutrients because the [Fe,Mn,Zn,Zn,Cu | B,Mg,Ca,K,P,N,S] fulcrum was driven on the positive side. Also, low Mg and low Ca misbalanced [Mg,Ca | K] while maintaining [Mg | Ca] in balance. The [N | S] balance was significantly dragged to the left by N excess. The 'Espada' [Fv | Nutrients] balance departed significantly from the TNs on the negative side, indicating overall nutrient shortage. Although balances related to micronutrients were within the TN range, 'Tommy' was largely misbalanced by K shortage and somehow by Mg shortage and, apparently, Ca excess. The addition of K and Mg fertilizers could re-establish the [Fv | Nutrients] balance. Most balances of 'Palmer' were within the TN balance ranges, except for [Mg,Ca | K], mostly due to K excess. This balance could be centered by adding Mg and Ca or omitting K additions, with some risk of misbalancing other balances.



Category	Fv	Fe	Mn	Zn	Cu	B	Mg	Ca	K	P	N	S
TN	953	0.0581	0.473	0.0198	0.0157	0.0261	1.90	19.9	9.2	0.98	13.1	1.36
TP	953	0.0641	0.418	0.0206	0.0272	0.0303	1.98	19.7	9.3	0.91	13.6	1.41
Tommy	954	0.0613	0.439	0.0192	0.0158	0.0251	1.87	21.1	8.0	0.95	12.1	1.29
Palmer	952	0.0600	0.501	0.0202	0.0181	0.0263	1.97	19.3	10.4	0.96	13.5	1.38
Haden	951	0.0447	0.455	0.0284	0.0239	0.0371	1.80	18.5	11.1	1.42	13.9	1.74
Espada	955	0.0480	0.328	0.0182	0.0106	0.0303	1.72	18.9	6.3	0.99	15.3	1.35

Figure 5. Pan balance design illustrating nutrient equilibrium in foliar tissues of mango cultivars. Concentrations in weighing pans are back-transformed *ilr* means of TN specimens.

4.5 Numerical biases

Because the *ilr* transformation leads to unbiased multivariate analysis, the deviation from linearity in the Mahalanobis distance of ordinary log transformed concentration data or the DRIS nutrient imbalance index is a measure of numerical bias similar to the Aitchison distance (Lovell, 2011). The ordinary log transformation as well as DRIS (using TN specimens as reference subpopulation to compute dual ratio norms for high yielders and the coefficient of variation) produced noisy, possibly leading to conflicting interpretations, diagnoses due to numerical biases (Figure 6).

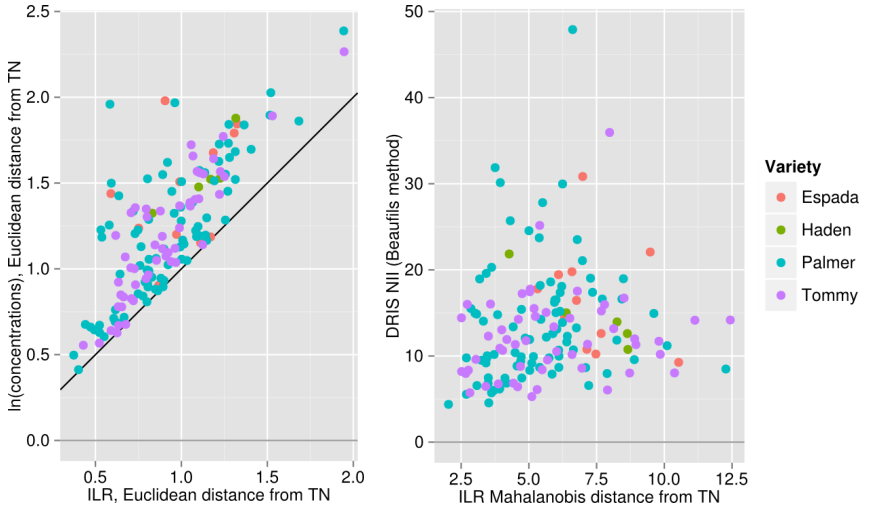


Figure 6. Numerical bias measured by discrepancy between isometric log ratio transformations of mango ionomes and (a) Euclidean distances computed ordinary log and DRIS nutrient imbalance index.

5 Discussion

5.1 Mango ionomes

Because mango varieties were developed essentially (90%) from the germplasm of *Mangifera indica* (Mukherjee, 1963), nutrient management of mango orchards are generally thought to be related to yield potential at species rather than variety level. Phenotypic plasticity is a phenomenon typical of domesticated

species that are most often bred for high productivity under relatively luxurious environments (Chapin, 1980, 1989). Due to different varieties being grown in different soil fertility categories, high variation in ionomes could not be interpreted as genotypic effect, because of potentially high phenotypic plasticity driven by differential nutrient supplies. There is often a large number of misclassified false negative specimens in fruit crop survey datasets not only due to small climatic variations and natural or pathological changes occurring in trees, but much more so due to biennal fruit bearing habits alternating between 'on-year' and 'off-year' (Monselise and Goldschmidt, 1982). However, proper pruning of the mango tree limited the effect of alternate bearing on fruit yield in the surveyed orchards.

5.2 Numerical biases

Filzmoser et al. (2009) argued that a log ratio transformation is similar to an ordinary log transformation only when a large filling value is used as denominator, because $\lim_{x_D \to 1} [alr(x) - ln(x)] = 0$. We showed that, for mango ionomes, numerical biases using ordinary log-transformed concentration data distorted the multivariate diagnosis despite large filling values. This result indicates that log ratio transformations are preferable to ordinary log transformations to avoid numerical biases when handling plant nutrient data.

Both CNCR and DRIS are not only numerically biased (Parent et al., 2012), but possibly lead to conflicting results when conducted separately (Silva et al., 2004; Blanco-Macias et al., 2009; Huang et al., 2012; Wairegi and van Asten, 2012). Despite apparent utility of dual ratios for diagnostic purposes, whether a nutrient level is too high, adequate or too low is impossible to determine (Walworth and Sumner, 1987; Marschner, 1995; Wilkinson et al., 2000). The pan balance approach connects nutrient balances and concentrations within a coherent, statistically unbiased, model where concentrations can assist in appreciating the results of statistical analyses on balances.

The analyst should be reminded that nutrient deficiency, sufficiency or excess of any nutrient can only be diagnosed in relation to other nutrients in a balance system. Only the balance can be tested statistically. The weighing pans facilitate interpreting the balances correctly. For example, the lower the [Mg | Ca] balance in TN specimens can be appreciated as a combination of lower Ca and higher Mg concentrations compared to TP specimens and this is ascertained looking at concentration values associated with the corresponding TN nutrient loads in weighing pans. Corrective measures involving one element may impact on all balances connected to it. This is why the effect of corrective measures on such complex system should be confirmed in many cases.

Using *ilrs* at fulcrums for unbiased diagnosis and nutrient concentrations in buckets to provide an appreciation of the results as relative shortage, adequacy or excess compared to TN barycentres, plant diagnosticians are informed at a glance and coherently on how concentration levels impact on nutrient balances. We thus suggest a change of paradigm from the traditionally combined and potentially conflicting CNCR-DRIS diagnoses of nutrient status based on Liebig's barrel to the stand-alone pan balance metaphor illustrated by a mobile-and fulcrums setup with weighing pans loaded with nutrient concentrations. Diagnosis of nutrient balances and related nutrient concentrations is therefore conducted coherently using a single setup without bias.

6 Conclusion

Using a Brazilian mango data set of crop productivity and plant and soil compositions, we addressed two typical problems when diagnosing the mineral nutrition of fruit crops: (1) genotype effect vs. phenotypic plasticity and (2) double-biased diagnosis with CNCR and DRIS conducted separately vs. coherent stand-alone balance-concentration setup. Ionomes of mango varieties were interpreted as phenotypic plasticity, and nutrient balance norms were developed at species level. Former diagnostic tools developed according to the 'Law of minimum' and illustrated by Liebig's barrel in agronomic studies should be replaced by modern tools of compositional data analysis because ionomes are vectors of compositional data. The pan balance metaphoric representation of *ilr* variables is a novel model that integrates statistical diagnosis of balances and qualitative evaluation of nutrient concentrations into a unified and coherent diagnosis that avoids numerical biases and conflicting interpretation of nutrient concentrations and ratios when conducted separately.

7 Acknowledgements

We acknowledge the financial support of the Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP, the Natural Sciences and Engineering Council of Canada (DG-2254 and CRDPJ 385199 – 09). We thank Brazilian mango growers and Canadian farm partners as follows: Cultures Dolbec Inc., St-Ubalde,

Québec, Canada; Groupe Gosselin FG Inc., Pont Rouge, Québec, Canada; Agriparmentier Inc. and Prochamps Inc., Notre-Dame-du-Bon-Conseil, Québec, Canada; Ferme Daniel Bolduc et Fils Inc., Péribonka, Québec, Canada.

8 References

Aitchison, J. 1986. The statistical analysis of compositional data. Chapman and Hall, London.

Aitchison, J. and M. Greenacre. 2002. Biplots of compositional data. J. Royal Stat. Soc. Series C (Appl. Stat.) 51(4):375-392.

Bacon-Shone, J. (2011). "A short history of compositional data analysis," in *Compositional data analysis: Theory and Applications*, ed. Pawlowsky-Glahn V., and Buccianti A., [NY: John Wiley and Sons], 3-11.

Baxter, I.R., Vitek, O., Lahner, B., Muthukumar, B., Borghi, M., Morrissey, J., Guerinot, M.L., and Salt, D.E. 2008. The leaf ionome as a multivariable system to detect a plant's physiological status. Proceedings of the National Academy of Science of United States of America, 105(33) : 12081-12086.

Benton Jones, J. Jr, B. Wolf and H.A. Mills. 1991. Plant analysis handbook. A practical sampling, preparation, analysis, and interpretation guide. Micro-Macro Publ., Athens, GA.

Bergmann, W. 1988. Ernährungs-störungen bei Kulturpflanzen. 2. Auflage. Gustav Fischer Verlag, Stuttgart, Germany.

Blanco-Macías, F., Magallanes-Quintanar, R., Valdez-Cepeda, R.D., Vázquez-Alvarado, R., Olivares-Sáenz, E., Gutiérrez-Ornelas, E., and Vidales-Contreras, J.A. 2009. Comparison between CND norms and boundary-line approach nutrient standards: Opuntia ficus indica L. case. Revista Chapingo Serie Horticultura 15(2): 217-223.

Chapin, S.F. III. 1989. Ecological aspects of plant nutrition. Adv. Plant Nutr. 3:161-191.

Chapin, S.F. III. 1980. The mineral nutrition of wild plants. Ann. Rev. Ecol. Syst. 11:233-260.

Chayes, F. 1960. On correlation between variables of constant sum. J. Geophys. Res. 65:4185-4193.

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

Egozcue, J.J., V. Pawlowsky-Glahn, G. Mateu-Figueras and C. Barceló-Vidal. 2003. Isometric log-ratio transformations for compositional data analysis. Math. Geol. 35:279-300.

Epstein, W., and Bloom, A. J. 2005. Mineral nutrition of plants: Principles and Perspectives. 2nd Edn. Sunderland MA: Sinauer Associates.

Filzmoser, P., K. Hron and C. Reimann. 2009. Univariate statistical analysis of environmental (compositional) data: Problems and possibilities. Sci. Total Environ. 407(23):6100-6108.

Filzmoser, P., and Hron, K. (2011). "Robust statistical analysis," in *Compositional data analysis: Theory and Applications*, ed. Pawlowsky-Glahn V., and Buccianti A., [NY: John Wiley and Sons], 57-72.

Filzmoser, P. and Gschwandtner, M. 2013. mvoutlier: Multivariate outlier detection based on robust methods in in R package version 1.9.9, http://CRAN.R-project.org/package=mvoutlier

Gang, L., Nunes, L., Wang, Y., Williams, P.N., Zheng, M., Zhang, Q. and Zhu, Y. 2013. Profiling the ionome of rice and its use in discriminating geographical origins at the regional scale, China. Journal of Evironmental Sciences, 25(1):144-154.

Hanley, J.E. 1988. The robustness of the "binormal" assumptions used in fitting ROC curves" Medical Decision Making 8: 197–203.

Han, W. X., Fang, J. Y., Reich, P. B., Woodward, F. I., and Wang, Z. H. (2011). Biogeography and variability of eleven mineral elements in plant leaves across gradients of climate, soil and plant functional type in China. *Ecol. Lett.* 14, 788–796.

Huang, H., Xiao Hu, C., Tan, Q., Hu, X., Sun X., and Bi, L. 2012. Effects of Fe–EDDHA application on iron chlorosis of citrus trees and comparison of evaluations on nutrient balance with three approaches. Scientia Hort. 146:137–142.

Ingestad T. 1987. New concepts on soil fertility and plant nutrition as illustrated by research on forest trees and stands. Geoderma 40 :237-252.

Jones, J.B. Jr. and V.W. Case. 1990. Sampling, handling, and analyzing plant tissue samples. p. 389-427 In R.L. Westerman (ed). Soil testing and plant analysis. 3rd ed., Soil Sci. Soc. Am. Book Ser. 3. SSSA, Madison, WI.Loladze, I., and Elser, J. J. 2011. The origins of the Redfield nitrogen-to-phosphorus ratio are in a homoeostatic protein-to-rRNA ratio. Ecol. Lett. 14, 244-250.

Lahner, B., Gong J., Mahmoudian, M., Smith, E. L., Abid, K. B., Rogers, E. E., Guerinot, M. L., Harper, J. F., Ward, J. M., McIntyre, L., Schroeder, J. I., and Salt, D. E. 2003. Genomic scale profiling of nutrient and trace elements in Arabidopsis thaliana. Nature Biotech. 21, 1215–1221.

Lovell, D., W. Müller, J. Taylor, A. Zwart and C. Helliwell. 2011. Proportions, percentages, ppm: do the molecular biosciences treat compositional data right? p. 193-207. In Pawlowsky-Glahn, V. and A. Buccianti, editors, Compositional Data Analysis: Theory and Applications, John Wiley and Sons, New-York.

Malavolta, E. 2006. Manual de nutrição de plantas. (In Portuguese.) Pav. Chimica, ESALQ and Ed. Agron. CERES, São Paulo, Brazil.

Marshner, H. 1995. Mineral nutritionof higher plants. Academic Press, NY, 674 pp.

Monselise, S.P. and E.E. Goldschmidt. 1982. Alternate bearing in fruit trees. Hort. Rev. 4:128-173.

Mukherjee, S.K. 1963. Citology and breeding of mango. Punjab Horticultural Journal, 3:107-115.

Nelson, L.A. And Anderson, R.L. 1977. Partitioning soil test – Crop response probability. In : Soil testing : Correlating and interpreting the analytical results, ASA special publication Number 29.

Parent, L. E. and Dafir, M. 1992. A theoretical concept of compositional nutrient diagnosis. J. Amer. Soc. Hort. Sci. 117:239-242.

Parent, S.- É., Parent, L. E. Rozane, D. E. Hernandes, A., Natale, W. 2012. Nutrient balance as paradigm of plant and soil chemometrics. Chapter 4 (32 pp.). In Issaka, R.N. (ed.). Soil Fertility, InTech Publ., http://www.intechopen.com/books/soil-fertility.

Parent, S.-É., Parent, L.E., Egozcue, J.J., Rozane, D.E., Hernandes, A., Lapointe, L., Hébert-Gentile, V., Naess, K., Marchand, S., Lafond, J., Mattos Jr, D. , Barlow, P. and Natale, W. The plant ionome revisited by the nutrient balance concept. Frontiers in Plant Science, 4: Article 39.

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. Proc. Roy. Soc, London LX:489-502.

Prekopcsák, Z. and Lemire, D. 2012. Time series classification by class-specific Mahalanobis distance measures. Advances in Data Analysis and Classification, 6(3):185-200.

Quaggio, J. A., Raij, B. van and Piza Junior, C. T. 1997. Frutíferas. In: Raij, B. van, Cantarella, H., Quaggio, J. A., Furlani, A. M. C. (Eds). Recomendações de adubação e calagem para o Estado de São Paulo. Instituto Agronômico/Fundação, Campinas, Brazil.

R Development Core Team. 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Version 2.15.2. http://www.R-project.org

Raij B. van, J.A. Quaggio, H. Cantarella, M.E. Ferreira, A.S. Lopes, and O.C. Bataglia. 1987. Análise química do solo para fins de fertilidade. Fundação Cargill, Campinas, Brazil.

Silva, G.G.C. da, Neves, J.C.L., Alvarez V.V.H., and Leite, F.P. 2004. Nutritional diagnosis for Euvcalypt by DRIS, M-DRIS , and CND. Sci. Agric. (Piracicaba, Braz.) 61(5):507-515.

Swets, J., 1988. Measuring the accuracy of diagnostic systems. Science, 240 :1285–1293.

Tanner, J. 1949. Fallacy of per-weight and per-surface area standards, and their relation to spurious correlation. J. Physiol. 2:1-15.

van den Boogaart, K. G., R. Tolosana-Delgado and M. Bren, M. 2013. compositions: Compositional Data Analysis in R package version 1.30-1, http://CRAN.R-project.org/package=compositions

Wadt, P.G.S. and D.J. Silva. 2010. Nutritional diagnosis accuracy of mango tree orchards through three DRIS formulas. Pesq. Agropec. Bras., Brasilia, 45(10):1180-1188.

Walworth, J.L. and M.E. Sumner. 1987. The Diagnosis and Recommendation Integrated System (DRIS). Adv. Soil Sci. 6:149-188.

Wairegi, L. and van Asten, P. 2012. Norms for multivariate diagnosis of nutrient imbalance in Arabica and Rosusta coffee in the East African Highlands. Expl Agric. 48 (3):448–460.

Wilkinson, S.R., D.L. Grunes and M.E. Sumner. 2000. Nutrient interactions in soil and plant nutrition. p. D89-D112. In Sumner, M.E., editor-in-chief, Handbook of soil science, CRC Press, Boca Raton FL.

The product space $\mathcal{T}$ (tools for compositional data with a total)

V. PAWLOWSKY-GLAHN[1], J. J. EGOZCUE[2], D. LOVELL[3]

[1] Dept. Informatics and Applied Mathematics, U. de Girona, Spain, vera.pawlowsky@udg.edu
[2] Dept. Applied Mathematics III, U. Politécnica de Catalunya, Barcelona, Spain
[3] CSIRO Mathematics, Informatics, and Statistics, Canberra, Australia

Abstract

The analysis of compositional data deals with relative information between parts. The total (abundances, mass, amount, ...) is in general not known, or not informative. Occasionally, interest lies in analysing a composition for which the total is known and of interest. Tools used in these cases are reviewed and analysed, in particular the relationship between the positive orthant of D-dimensional real space, $\mathbb{R}_+^D$, the product space $\mathbb{R}_+ \times \mathcal{S}^D = \mathcal{T}$, and their Euclidean space structures. Real data about total abundances of phytoplankton in an Australian river motivated the present study and are used for illustration.

1 Introduction

Compositional data analysis usually starts with data presented as vectors in the positive orthant of D-dimensional real space, $\mathbb{R}_+^D$. If interest lies solely in comparing data independently of their size, standard practice is to project them in the simplex $\mathcal{S}^D$, the subset of $\mathbb{R}_+^D$ that results typically from constraining the data to sum to a constant κ, herein assumed to be $\kappa = 1$. The D-part simplex, $\mathcal{S}^D$, is defined as the set of real vectors with positive components adding to the constant $\kappa = 1$. The projection is attained applying the *closure operation* $\mathcal{C}$ (Aitchison, 1986), which consists in dividing each component by the sum of all components. This approach enables compositional analysis, but ignores information about total abundance. To analyse the whole vector, including information about total abundance, two alternatives are usual in practice. One consists in considering $\ln(\mathbf{x})$, where the logarithm applies componentwise. The other one involves projecting the composition into $\mathcal{S}^D$ and considering the total sum as an additional variable. To perform an analysis of the obtained vectors using standard methods, it is necessary to go a step further and express them as coordinates in real space.

Here, we examine which are suitable Euclidean structures of the support spaces under consideration, with the assumption that components are strictly positive. This assumption implies that zeros are not considered as possible values, although samples frequently include them. To our understanding, zeros require a special treatment, similar to that used for compositional data, and are beyond the scope of this initial paper.

In what follows, vectors in $\mathbb{R}_+^D$ will be denoted by boldface, lowercase letters, e.g. $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$, and their components by lowercase letters with an index, i.e. $\mathbf{x} = [x_1, x_2, \ldots, x_D]$, where the square brackets indicate row vectors. Vectors in the product space $\mathbb{R}_+ \times \mathcal{S}^D$, as well as their components, will be denoted with a tilde, i.e. $\tilde{\mathbf{x}}$, $\tilde{\mathbf{y}}$, $\tilde{\mathbf{z}}$, where $\tilde{\mathbf{x}} = [t_x, \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_D]$, and t_x denotes some total of interest.

2 Space structure of the positive orthant $\mathbb{R}_+^D$

Consider a vector with D strictly positive components, $\mathbf{x} \in \mathbb{R}_+^D$. In $\mathbb{R}_+^D$, the logarithmic transformation applied to each component induces a Euclidean space structure over $\mathbb{R}$. It is a straightforward extension to $\mathbb{R}_+^D$ of the Euclidean structure of the positive real line, $\mathbb{R}_+$ (Pawlowsky-Glahn and Egozcue, 2001). The Euclidean structure demands an Abelian group operation, an external multiplication, and an inner product.

Definition 2.1 (Euclidean structure of $\mathbb{R}_+^D$) *The Abelian inner group operation is called* plus-perturbation*, the external multiplication is called* plus-powering*. For $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D$ and $\alpha \in \mathbb{R}$, they are*

defined as

$$\mathbf{x} \oplus_+ \mathbf{y} = [x_1 \cdot y_1, x_2 \cdot y_2, \ldots, x_D \cdot y_D] \ , \ \alpha \odot_+ \mathbf{x} = [x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha] \ , \tag{1}$$

respectively. The inner product in $\mathbb{R}_+^D$*, for* $\mathbf{x}, \mathbf{y} \in \mathbb{R}_+^D$*, is called* plus-inner-product *and is defined as*

$$\langle \mathbf{x}, \mathbf{y} \rangle_+ = \langle \ln \mathbf{x}, \ln \mathbf{y} \rangle \ , \tag{2}$$

where $\langle \, , \, \rangle$ *stands for the usual, Euclidean, inner product in* $\mathbb{R}^D$ *and the logarithm,* ln*, applies componentwise.*

Plus-perturbation is an Abelian group operation because the product of positive numbers satisfies all the required axioms in each component (associativity and commutativity, existence of neutral and inverse element). The neutral element (identity) is $\mathbf{n}_+ = [1, 1, \ldots, 1]$; the inverse element of $\mathbf{x}$ is $\ominus_+ \mathbf{x} = [1/x_1, 1/x_2, \ldots, 1/x_D]$. The operation $\oplus_+$ is called *plus-perturbation*, by analogy to the group operation in the simplex, called *perturbation* (Aitchison, 1986; Pawlowsky-Glahn and Egozcue, 2001).

Plus-powering is distributive with respect to the vector group operation, $\alpha \odot_+ (\mathbf{x} \oplus_+ \mathbf{y}) = (\alpha \odot_+ \mathbf{x}) \oplus_+ (\alpha \odot_+ \mathbf{y})$; distributive with respect to field addition, $(\alpha + \beta) \odot_+ \mathbf{x} = (\alpha \odot_+ \mathbf{x}) \oplus_+ (\beta \odot_+ \mathbf{x})$; compatible with field multiplication, $\alpha \odot_+ (\beta \odot_+ \mathbf{x}) = (\alpha \cdot \beta) \odot_+ \mathbf{x}$; and has an identity element, $1 \odot_+ \mathbf{x} = \mathbf{x}$, where 1 denotes the multiplicative identity in $\mathbb{R}$.

With the above operations and inner product, $\mathbb{R}_+^D$ is a D-dimensional real Euclidean vector space. Furthermore, the associated distance and norm are:

$$d_+ (\mathbf{x}, \mathbf{y}) = d (\ln \mathbf{x}, \ln \mathbf{y}) \ ; \quad \|\mathbf{x}\|_+ = \|\ln \mathbf{x}\| \ , \tag{3}$$

where d and $\| \cdot \|$ stand for the Euclidean distance and norm in $\mathbb{R}^D$.

Note that the above definitions correspond to the standard practice of analysing abundances taking logarithms and looking at deviations using the difference of logarithms. However, $\mathbb{R}_+^D$ admits alternative Euclidean structures with a different definition of the group operation $\oplus_+$ and also different definition of the metrics, as shown below.

The above definitions and properties raises two questions, (1) when it is appropriate to analyse a composition for which total abundances have been measured in $\mathbb{R}_+^D$ with the above structure, and (2) what is the difference to analysing this data as a composition and treat the *total* as an external variable. To explore these questions and see where the two approaches differ, we need to consider the product space $\mathbb{R}_+ \times \mathcal{S}^D$.

3 Space structure of $\mathbb{R}_+ \times \mathcal{S}^D = \mathcal{T}$

Consider a vector with D strictly positive components, $\mathbf{x} \in \mathbb{R}_+^D$. Applying the closure operation yields a D-part composition $\mathcal{C}(\mathbf{x}) \in \mathcal{S}^D$. An associated positive value $t(\mathbf{x}) \in \mathbb{R}_+$ can be defined with an appropriate function. This function $t(\cdot)$ can be the total sum (abundance, mass, ...), the product, the arithmetic mean, the geometric mean, a single component or any other value related to the problem. Some particular cases will be discussed below.

The vector $\tilde{\mathbf{x}} = [t(\mathbf{x}), \mathcal{C}(\mathbf{x})] = [t_x, \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_D]$ is an element of $\mathbb{R}_+ \times \mathcal{S}^D$, the product space of the sample spaces (the sets of possible values) of $t(\mathbf{x})$ and $\mathcal{C}(\mathbf{x})$. For $D = 3$, this product space can be visualized as a 3-part simplex (or ternary diagram) with an orthogonally attached positive real line. In what follows, it will be denoted by $\mathcal{T}$. As before, to define a Euclidean space structure in $\mathcal{T}$, an Abelian inner group operation, an external multiplication, and an inner product are needed.

Definition 3.1 (Euclidean structure of $\mathbb{R}_+ \times \mathcal{S}^D = \mathcal{T}$) *The Abelian inner group operation and the external multiplication in* $\mathcal{T}$*, are called* $\mathcal{T}$-perturbation *and* $\mathcal{T}$-powering*. For* $\tilde{\mathbf{x}}, \tilde{\mathbf{y}} \in \mathcal{T}$*, and* $\alpha \in \mathbb{R}$*, they are respectively defined as*

$$\tilde{\mathbf{x}} \oplus_T \tilde{\mathbf{y}} = [t(\mathbf{x}) \oplus_+ t(\mathbf{y}), \mathbf{x} \oplus_a \mathbf{y}] = [t_x \cdot t_y, \mathcal{C}(\tilde{x}_1 \tilde{y}_1, \tilde{x}_2 \tilde{y}_2, \ldots, \tilde{x}_D \tilde{y}_D)] \ , \tag{4}$$

$$\alpha \odot_T \tilde{\mathbf{x}} = [\alpha \odot_+ t(\mathbf{x}), \alpha \odot_a \mathbf{x}] = [t_x^\alpha, \mathcal{C}(\tilde{x}_1^\alpha, \tilde{x}_2^\alpha, \ldots, \tilde{x}_D^\alpha)] \ , \tag{5}$$

where $\oplus_+$, $\odot_+$ *stand for perturbation and powering in* $\mathbb{R}_+$, *and* $\oplus_a$ *and* $\odot_a$ *for perturbation and powering in* $\mathcal{S}^D$ *(Pawlowsky-Glahn and Egozcue, 2001). The inner product in* $\mathcal{T}$ *is called* $\mathcal{T}$-inner-product *and is defined, for* $\tilde{\mathbf{x}}, \tilde{\mathbf{y}} \in \mathcal{T}$, *as*

$$\langle \tilde{\mathbf{x}}, \tilde{\mathbf{y}} \rangle_T = \langle t_x, t_y \rangle_+ + \langle \mathcal{C}(\mathbf{x}), \mathcal{C}(\mathbf{y}) \rangle_a \,, \tag{6}$$

where $\langle \, , \rangle_+$ *stands for the inner product in* $\mathbb{R}_+$, *and* $\langle \, , \rangle_a$ *stands for the Aitchison inner product in* $\mathcal{S}^D$ *(Pawlowsky-Glahn and Egozcue, 2001).*

As $(\oplus_+, \odot_+, \langle \cdot, \cdot \rangle_+)$ and $(\oplus_a, \odot_a, \langle \cdot, \cdot \rangle_a)$ define the operations and metrics in, respectively, $\mathbb{R}_+$ and $\mathcal{S}^D$, it follows that the same holds for $\oplus_T$, $\odot_T$ and $\langle \cdot, \cdot \rangle_T$ in $\mathcal{T}$. Note that the same symbols are used to denote operations in $\mathbb{R}_+^D$ and $\mathbb{R}_+$, as they are identical, except that one operates on a vector, while the other operates on a single variable.

Theorem 3.1 (Euclidean vector space) *The product space* $\mathcal{T} = \mathbb{R}_+ \times \mathcal{S}^D$, *with* $\mathcal{T}$*-perturbation* $(\oplus_T)$, $\mathcal{T}$*-powering* $(\odot_T)$, *and* $\mathcal{T}$*-inner product* $(\langle \cdot, \cdot \rangle_T)$ *is a* D*-dimensional Euclidean vector space on* $\mathbb{R}$.

Definition 3.1 implies that the square distance in $\mathcal{T}$ is

$$d_T^2(\tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = d_+^2\left(t(\mathbf{x}), t(\mathbf{y})\right) + d_a^2(\mathcal{C}(\mathbf{x}), \mathcal{C}(\mathbf{y})) = \ln^2 \frac{t_x}{t_y} + d_a^2(\mathcal{C}(\mathbf{x}), \mathcal{C}(\mathbf{y})) \,. \tag{7}$$

4 Hypotheses and compatibility

Interest now centres on vectors $\mathbf{x} \in \mathbb{R}_+^D$ for which $\mathcal{C}(\mathbf{x})$ is assumed to obey the Aitchison geometry of $\mathcal{S}^D$ and the principles of compositional data analysis, so that the statistics or calculus performed on $\mathcal{C}(\mathbf{x})$ are compatible with those performed on $\mathcal{T} = \mathbb{R}_+ \times \mathcal{S}^D$. The Euclidean structure of $\mathbb{R}_+$ is also assumed, i.e. the operations $\oplus_+$ and $\odot_+$ work properly for $t(\cdot)$.

The space $\mathbb{R}_+^D$ is involved in this compatibility. Certainly, the perturbation in $\mathbb{R}_+^D$, $\oplus_+$, must satisfy $\mathcal{C}(\mathbf{x} \oplus_+ \mathbf{y}) = \mathcal{C}(\mathbf{x}) \oplus_a \mathcal{C}(\mathbf{y})$, as a consequence of the scale invariance principle of compositional data analysis. Then, the compatibility is focused on the transformation $h : \mathbb{R}_+^D \to \mathcal{T}$, whose expression must be $h(\mathbf{x}) = [t(\mathbf{x}), \mathcal{C}(\mathbf{x})]$. Specifically, the conditions are on the total function $t(\mathbf{x})$. The map h must be one-to-one, otherwise some information is lost when applying h or h^{-1}. Since $\mathbf{x} = \left(\sum_{i=1}^D x_i\right) \cdot \mathcal{C}(\mathbf{x})$, the total function $t(\mathbf{x})$ has to be related to the sum of the components, to allow the reconstruction of $\mathbf{x}$ from the composition and the total. Therefore, the first compatibility condition is that $h : \mathbb{R}_+^D \to \mathcal{T}$ is a one-to-one mapping.

The second compatibility condition on h, or alternatively on $t(\cdot)$, is the preservation of the vector space properties in $\mathbb{R}_+^D$ and $\mathcal{T}$. This condition can be written as

$$h(\mathbf{x} \oplus_+ \mathbf{y}) = h(\mathbf{x}) \oplus_T h(\mathbf{y}) \quad , \quad h(\alpha \odot_+ \mathbf{x}) = \alpha \odot_T h(\mathbf{x}) \,. \tag{8}$$

Expressing the conditions (8) for the component involving the total t gives

$$t(\mathbf{x} \oplus_+ \mathbf{y}) = t(\mathbf{x}) \cdot t(\mathbf{y}) \quad , \quad t(\alpha \odot_+ \mathbf{x}) = [t(\mathbf{x})]^\alpha \,.$$

The metric in $\mathcal{T}$ is implied by the characteristics of product space reflected in Eq. (6) and (7). By construction, distances in $\mathcal{T}$ dominate distances in $\mathcal{S}^D$ (Eq. 7) and therefore, for any $\tilde{\mathbf{x}}$ $\tilde{\mathbf{y}}$ in $\mathcal{T}$, $d_T(\mathbf{x}, \mathbf{y}) \geq d_a(\mathcal{C}(\mathbf{x}), \mathcal{C}(\mathbf{y}))$. However, when the total $t(\mathbf{x})$ is specified, the map h^{-1} induces a metric in $\mathbb{R}_+^D$ that can differ from the standard one (Eq. 2).

As shown in Lovell et al. (2011), for the distance d_+ it holds

$$d_+^2(\mathbf{x}, \mathbf{y}) = d_a^2(\mathcal{C}(\mathbf{x}), \mathcal{C}(\mathbf{y})) + D \ln^2 \left[\frac{\mathrm{g_m}(\mathbf{x})}{\mathrm{g_m}(\mathbf{y})}\right] \,, \tag{9}$$

which should be compared to (7). This suggests that equality will be obtained for $t(\mathbf{x}) = \mathrm{g_m}(\mathbf{x})$, or some power thereof, as discussed below.

4.1 The *total* as a power of the product of components

Consider $t_\delta : \mathbb{R}_+^D \to \mathbb{R}_+$ to be the product of the components of $\mathbf{x} \in \mathbb{R}_+^D$ to the power δ, i.e. $t_\delta(\mathbf{x}) = \prod_{i=1}^D x_i^\delta$. The following properties hold:

Proposition 4.1 *Let be* $h : \mathbb{R}_+^D \to \mathcal{T}$ *such that*

$$h(\mathbf{x}) = [t_\delta(\mathbf{x}), \mathcal{C}(\mathbf{x})] = \left[\prod_{i=1}^D x_i^\delta, \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_D\right] .$$

(a) h *is one-to-one; (b)* h *is compatible with* $\oplus_+$ *and* $\odot_+$*, i.e. for* $\alpha \in \mathbb{R}$ $h((\alpha \odot_+ \mathbf{x}) \oplus_+ \mathbf{y}) = (\alpha \odot_T h(\mathbf{x})) \oplus_T h(\mathbf{y})$.

Proof

(a) $\boxed{\Rightarrow}$ Given $\mathbf{x} \in \mathbb{R}_+^D$, by definition $h(\mathbf{x}) \in \mathcal{T}$.
$\boxed{\Leftarrow}$ Herein, t_x^δ denotes $t_\delta(\mathbf{x})$ for simplicity. Given $\tilde{\mathbf{x}} = \left[t_x^\delta, \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_D\right] \in \mathcal{T}$, and given that

$$\prod_{i=1}^D \tilde{x}_i = \prod_{i=1}^D \frac{x_i}{\sum_{j=1}^D x_j} = \frac{\prod_{i=1}^D x_i}{\left[\sum_{j=1}^D x_j\right]^D} = \frac{\left[\prod_{i=1}^D x_i^\delta\right]^{1/\delta}}{\left[\sum_{j=1}^D x_j\right]^D} = \frac{(t_x^\delta)^{1/\delta}}{\left[\sum_{j=1}^D x_j\right]^D} ,$$

it holds that

$$\sum_{j=1}^D x_j = \frac{(t_x^\delta)^{1/(D\delta)}}{\prod_{i=1}^D \tilde{x}_i^{1/D}} = \frac{(t_x)^{1/D}}{\left(\prod_{i=1}^D \tilde{x}_i\right)^{1/D}} = \frac{(t_x)^{1/D}}{\mathrm{g_m}(\mathcal{C}(\mathbf{x}))} ,$$

and therefore

$$\mathbf{x} = h^{-1}(\tilde{\mathbf{x}}) = \left(\sum_{j=1}^D x_j\right) [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_D] = \frac{(t_x)^{1/D}}{\mathrm{g_m}(\mathcal{C}(\mathbf{x}))} [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_D] .$$

(b) Given that $t_{xy}^\delta = \left[\prod_{i=1}^D x_i y_i\right]^\delta = \prod_{i=1}^D x_i^\delta \prod_{i=1}^D y_i^\delta = t_x^\delta t_y^\delta$, it holds $h(\mathbf{x} \oplus_+ \mathbf{y}) = \left[t_{xy}^\delta, \mathcal{C}(\mathbf{x} \oplus_+ \mathbf{y})\right] = \left[t_x^\delta t_y^\delta, \mathcal{C}(\mathbf{x}) \oplus_a \mathcal{C}(\mathbf{y})\right] = h(\mathbf{x}) \oplus_T h(\mathbf{y})$. Similarly, for powering, $h(\alpha \odot_+ \mathbf{x}) = \left[t_x^{\alpha\delta}, \mathcal{C}(\mathbf{x}^\alpha)\right] = \left[(\alpha \odot_+ t_x^\delta), \mathcal{C}(\alpha \odot_+ \mathbf{x})\right] = \alpha \odot_T h(\mathbf{x})$.

∎

For $\delta = 1/\sqrt{D}$ the compatibility of operations is guaranteed by Proposition 4.1. Additionally, the mapping h becomes an isometry as the contribution of the total to the square distance is

$$\ln^2\left[\frac{t^{1/\sqrt{D}}(\mathbf{x})}{t^{1/\sqrt{D}}(\mathbf{y})}\right] = D \cdot \ln^2\left[\frac{\mathrm{g_m}(\mathbf{x})}{\mathrm{g_m}(\mathbf{y})}\right] ,$$

to be compared with Eq. (9). When the total is $t^{1/\sqrt{D}}(\mathbf{x})$, the product space $\mathcal{T}$ is denoted $\mathcal{T}_p$ and all operations and metric symbols are subscripted with p. With this notation, the following proposition holds.

Proposition 4.2 *If* $t_p(\mathbf{x}) = \prod_{i=1}^D x_i^{1/\sqrt{D}}$*, the Euclidean spaces* $\mathbb{R}_+^D$ *and* $\mathcal{T}_p$ *are isometric.*

4.2 The *total* as a sum

Consider $t : \mathbb{R}_+^D \to \mathbb{R}_+$ to be the sum of the components of $\mathbf{x}$, i.e. $t(\mathbf{x}) = t_s(\mathbf{x}) = \sum_{i=1}^D x_i$. Under such an assumption the product space $\mathcal{T}$ is denoted $\mathcal{T}_s$ and all operations and metric symbols are subscripted with s. The following properties hold:

Proposition 4.3 *Let $h_s : \mathbb{R}_+^D \to \mathcal{T}_s$ be such that*

$$h_s(\mathbf{x}) = [t_s(\mathbf{x}), \mathcal{C}(\mathbf{x})] = \left[\sum_{i=1}^D x_i, \mathcal{C}(\mathbf{x})\right] .$$

Then,(a) h_s is one-to-one; (b) h_s is not compatible with $\oplus_+$, $\odot_+$ and $\oplus_s$, $\odot_s$, i.e. $h_s((\alpha \odot_+ \mathbf{x}) \oplus_+ \mathbf{y}) \neq (\alpha \odot_s h_s(\mathbf{x})) \oplus_s h_s(\mathbf{y})$.

Proof

(a) $\boxed{\Rightarrow}$ Given $\mathbf{x} \in \mathbb{R}_+^D$, by definition $h_s(\mathbf{x}) \in \mathcal{T}$.
$\boxed{\Leftarrow}$ Given $\tilde{\mathbf{x}} = [t_s(\mathbf{x}), \mathcal{C}(\mathbf{x})] \in \mathcal{T}_s$, $\mathbf{x} = t_s(\mathbf{x})\mathcal{C}(\mathbf{x})$.

(b) The total of $\mathbf{x} \oplus_+ \mathbf{y}$ is $t_s(\mathbf{x} \oplus_+ \mathbf{y}) = \sum_{i=1}^D x_i y_i$, while $\tilde{\mathbf{x}} \oplus_s \tilde{\mathbf{y}}$ gives $t_s(\mathbf{x}) \oplus_+ t_s(\mathbf{y}) = \left[\sum_{i=1}^D x_i\right] \left[\sum_{i=1}^D y_i\right]$.
As $t_s(\mathbf{x} \oplus_+ \mathbf{y}) = \sum_{i=1}^D x_i y_i \neq \left[\sum_{i=1}^D x_i\right] \left[\sum_{i=1}^D y_i\right]$, it yields

$$h_s(\mathbf{x} \oplus_+ \mathbf{y}) = [t_s(\mathbf{x} \oplus_+ \mathbf{y}), \mathcal{C}(\mathbf{x} \oplus_+ \mathbf{y})] \neq [t_s(\mathbf{x}) \oplus_+ t_s(\mathbf{y}), \mathcal{C}(\mathbf{x}) \oplus_a \mathcal{C}(\mathbf{y})] = h_s(\mathbf{x}) \oplus_T h(\mathbf{y}) .$$

Similarly,

$$h_s(\alpha \odot_+ \mathbf{x}) = \left[\left(\sum_i x_i^\alpha\right), \mathcal{C}(\alpha \odot_+ \mathbf{x})\right] \neq [\alpha \odot_+ t_s(\mathbf{x}), \mathcal{C}(\alpha \odot_+ \mathbf{x})] = \alpha \odot_T h_s(\mathbf{x}) .$$

∎

However, the function h_s is one-to-one between $\mathbb{R}_+^D$ and $\mathcal{T}_s$ and, therefore, a Euclidean structure exists in $\mathbb{R}_+^D$ which is isometric to that one assumed in $\mathcal{T}_s$. The definitions of vector space operations and the metric are straightforward:

$$\mathbf{x} \oplus_{+s} \mathbf{y} = h_s^{-1}(\tilde{\mathbf{x}}) \oplus_s h_s^{-1}(\tilde{\mathbf{y}}) \ , \ \alpha \odot_{+s} \mathbf{x} = \alpha \odot_s h_s^{-1}(\tilde{\mathbf{x}}) \ ,$$

$$d_{+s}^2(\mathbf{x}, \mathbf{y}) = d_s^2(h_s(\mathbf{x}, \mathbf{y})) \ ,$$

where $\oplus_{+s}$ and $\odot_{+s}$ are the new operations in $\mathbb{R}_+^D$, compatible with the operations in $\mathcal{T}_s$ when the sum of elements is taken as the total. After some algebra the operations are

$$\mathbf{x} \oplus_{+s} \mathbf{y} = \left[\ldots, \frac{(\sum_k x_k)\,(\sum_\ell y_\ell)}{\sum_j x_j y_j}\, x_i y_i, \ldots\right] \ , \ \alpha \odot_{+s} \mathbf{x} = \left[\ldots, \frac{(\sum_k x_k)^\alpha}{\sum_j x_j^\alpha}\, x_i^\alpha, \ldots\right] \ ,$$

and the squared distance is

$$d_{+s}^2(\mathbf{x}, \mathbf{y}) = \ln^2\left(\frac{\sum_\ell x_\ell}{\sum_k y_k}\right) + d_a^2(\mathcal{C}(\mathbf{x}), \mathcal{C}(\mathbf{y})) \ .$$

5 Statistical consequences

In practical situations, one has to decide which is the relevant total function $t(\cdot)$ to use. When the total of interest is the sum of all components of a random vector $\mathbf{X}$ in $\mathbb{R}_+^D$, i.e. $t_s(\mathbf{X}) = \sum_{j=1}^D X_j$, the analysis to be conducted concerns random elements of $\mathcal{T}_s$ such as $h_s(\mathbf{X}) = [t_s(\mathbf{X}), \mathcal{C}(\mathbf{X})]$, where the subscript s indicates that the total t_s is considered. On the other hand, the analysis in $\mathbb{R}_+^D$ is often

performed. This analysis is isometric to that performed on $\mathcal{T}_p$, using the total $t_p(\mathbf{X}) = (\prod_i X_i)^{1/\sqrt{D}}$, and $h_p(\mathbf{X}) = [t_p(\mathbf{X}), \mathcal{C}(\mathbf{X})]$, with the subscript p indicating the product total.

The mean values or centres in $\mathbb{R}^D_+$, $\mathcal{T}_p$, $\mathcal{T}_s$ can be defined following the line introduced by Fréchet (1948) as the vectors which minimise metric variance (Pawlowsky-Glahn and Egozcue, 2001). After expressing the vectors in orthogonal coordinates of the corresponding space, they are

$$\begin{aligned}
\mathrm{Cen}_+(\mathbf{X}) &= \exp(\ \mathrm{E}[\ln \mathbf{X}]\ )\ , \\
\mathrm{Cen}_p(h_p(\mathbf{X})) &= [\ \exp\left(\mathrm{E}\left(\ln(t_p(\mathbf{X}))\right)\right)\ ,\ \mathrm{Cen}_a(\mathcal{C}(\mathbf{X}))\ ]\ , \\
\mathrm{Cen}_s(h_s(\mathbf{X})) &= [\ \exp\left(\mathrm{E}\left(\ln(t_s(\mathbf{X}))\right)\right)\ ,\ \mathrm{Cen}_a(\mathcal{C}(\mathbf{X}))\ ]\ ,
\end{aligned}$$

where E is the expectation for real random vectors and Cen_a is the centre of a composition using the Aitchison geometry. More explicitly,

$$\mathrm{E}(\ln(t_p(\mathbf{X}))) = \frac{1}{\sqrt{D}} \sum_{i=1}^{D} \mathrm{E}(\ln X_i) \quad , \quad \mathrm{E}(\ln(t_s(\mathbf{X}))) = \mathrm{E}\left(\ln\left(\sum_{i=1}^{D} X_i\right)\right)\ ,$$

which shows that $\mathrm{E}(\ln(t_p(\mathbf{X}))) \neq \mathrm{E}(\ln(t_s(\mathbf{X})))$ and that $\mathrm{Cen}_p(h_p(\mathbf{X})) \neq \mathrm{Cen}_s(h_s(\mathbf{X}))$. On the other hand, the isometry between $\mathbb{R}^D_+$ and $\mathcal{T}_p$ implies $h_p(\mathrm{Cen}_+(\mathbf{X})) = \mathrm{Cen}_p(h_p(\mathbf{X}))$. Given the incompatibility stated in Proposition 4.3, substantial differences between the statistical analysis carried out in $\mathbb{R}^D_+$ or, equivalently in $\mathcal{T}_p$, and $\mathcal{T}_s$ are expected.

Metric variances also follow the same rule

$$\mathrm{Mvar}_+(\mathbf{X}) = \mathrm{Mvar}_p(h_p((\mathbf{X}))) \neq \mathrm{Mvar}_s(h_s((\mathbf{X})))\ .$$

The difference between Mvar_p and Mvar_s can be made explicit using orthonormal coordinates in $\mathcal{T}$. The orthogonal coordinates in the simplex are taken equal in $\mathcal{T}_p$ and $\mathcal{T}_s$; the only additional coordinates are the logarithms of the totals. Therefore, as the metric variance is obtained adding up the variances of the coordinates (Pawlowsky-Glahn and Egozcue, 2001), it yields

$$\mathrm{Mvar}_p(h_p((\mathbf{X}))) - \mathrm{Mvar}_s(h_s((\mathbf{X}))) = \mathrm{Var}[\ln(t_p(\mathbf{X}))] - \mathrm{Var}[\ln(t_s(\mathbf{X}))]\ .$$

Centres and metric variances, in $\mathbb{R}^D_+$, $\mathcal{T}_p$, and $\mathcal{T}_s$, have the standard properties of means and variances. Particularly, if $\mathbf{X}$ is perturbed, its center is equally perturbed and the metric variance remains equal.

Another important issue is how centres change when units are changed in $\mathbb{R}^D_+$. Assume that the random vector $\mathbf{X} \in \mathbb{R}^D_+$ is multiplied by a positive constant a, equivalent to a change of units. Let $\mathbf{Y} = [aX_1, aX_2, \ldots, aX_D]$. This is equivalent to the perturbation

$$\mathbf{Y} = \mathbf{a} \oplus_+ \mathbf{X} \quad , \quad \mathbf{a} = [a, a, \ldots, a] \in \mathbb{R}^D_+\ .$$

Consequently, $\mathrm{Cen}_+(\mathbf{Y}) = \mathbf{a} \oplus_+ \mathrm{Cen}_+(\mathbf{X})$ and $\mathrm{Mvar}_+(\mathbf{Y}) = \mathrm{Mvar}_+(\mathbf{X})$. Equivalently in $\mathcal{T}_p$,

$$h_p(\mathbf{Y}) = [\ a^{\sqrt{D}}, \mathcal{C}(\mathbf{a})\ ] \oplus_s [\ t_p(\mathbf{X}), \mathcal{C}(\mathbf{X})\ ]\ ,$$

which leads to $\mathrm{Cen}_p(h_p(\mathbf{Y})) = h_p(\mathbf{a}) \oplus_p \mathrm{Cen}_p(h_p(\mathbf{X}))$ and $\mathrm{Mvar}_p(h_p(\mathbf{Y})) = \mathrm{Mvar}_p(h_p(\mathbf{X}))$.

In $\mathcal{T}_s$, these properties change a bit, as the number of components does not appear in the total of the perturbation

$$h_s(\mathbf{Y}) = [\ a, \mathcal{C}(\mathbf{a}/D)\ ] \oplus_s h_s(\mathbf{X})\ ,$$

and, therefore, $\mathrm{Cen}_s(h_s(\mathbf{Y})) = h_s(\mathbf{a}/D) \oplus_s \mathrm{Cen}_s(\mathbf{X})$ and $\mathrm{Mvar}_s(h_s(\mathbf{Y})) = \mathrm{Mvar}_s(h_s(\mathbf{X}))$.

Summarising, all multivariate techniques involving centres, metric variances, distances, or orthogonal projections, can give different results depending on the characteristics of the sample. The key point is the hypotheses on the sample space and its structure. If ratios of components of the data are relevant (i.e. there is compositional information, and the interesting total is the sum of the components), then the results of an analysis in $\mathcal{T}_s$ are different from those obtained using the geometry on coordinates in $\mathbb{R}^D_+$ introduced in Section 2, called hereafter *log-geometry*. The compatibility between the Euclidean geometry of $\mathbb{R}^D_+$ is only attained with $\mathcal{T}_p$ when the total is defined as the product of components to the power $1/\sqrt{D}$, not a very intuitive choice. Other options for the total, e.g. the geometric mean of the components, can make compatible the vector space structure but they do not provide an isometry.

6 Phytoplankton abundances: an example

Phytoplankton abundances were measured over a period of 14 years in an Australian river. Attention was centred in sampling on seasons 1 and 2 (from January to June), in which the sampling rate was approximately constant. Interest was in the evolution of the total abundance in time, and its possible relation to taxa present. After removal of incomplete compositions, available data consist of 173 samples of cyanobacteria and algae in time. The cyanobacteria are a subset of more toxic phytoplankton, often described as blue-green algae. For illustration, the following taxa where selected: *Anabaena* (ana), *Aphanizomenon* (aph) and *Other Cyanophyceae* (ocy) of the group of *Cyanobacteria*, and *Actinastrum* (act), *Cryptophyceae* (cry), *Ankistrodemus* (ank), *Aulacoseira distans* (aud) and *Aulacoseira granulata* (aug) of the group of *Algae*.

The exploratory analysis suggested that there could be two different regimes, before 1998 (labelled B, 58 samples), and after the beginning of 1998 (labelled A, 115 samples). In the whole data set there were substantial changes on the sampling density in time and the presence of non available (NA) data. The NA's for some taxa cannot clearly be ascribed to missing, non detected, or not present. These irregularities in the data collection were less apparent in the seasons 1 and 2 and therefore, they were selected to check whether there is a statistical difference in the centre values of samples B and A. Differences in composition are of interest, but also the total abundance is relevant in the analysis. There is a suspicious increment of the total abundance after 1998.

A first inspection of the data, using a blind SBP (Sequential Binary Partition) and the associated ilr transformation, shown in Figure 1, evidenced a different behaviour of the samples taken before

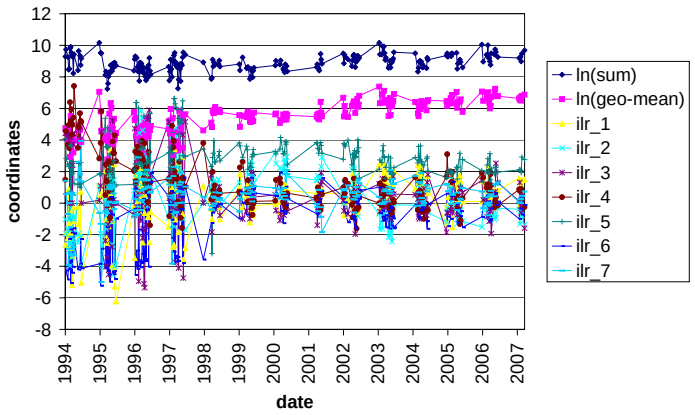


Figure 1: Preliminary exploration of Phytoplankton data: ilr transformed composition, log-sum and log-geometric mean as functions of time.

(B) and after (A) January 1, 1998. The following analysis treats the time periods B and A as two different samples. They are analysed using the $\mathbb{R}^D_+$, ($D = 8$), $\mathcal{T}_p$ ($t_p = \prod x_i^{1/\sqrt{D}}$) and $\mathcal{T}_s$ ($t_s = \sum x_i$) approaches, thus showing their equivalences and differences.

A first step is computing the centres for samples B and A, and considering all the samples belonging to the same sample, labelled BA. Table 1 shows the centres for A, B, and the joint sample BA. The upper part of the table is the centre computed in $\mathbb{R}^D_+$ and given in abundances per liter. Under the labels t_s and t_p, the totals corresponding to Cen_+ are specified. The lower part of the table shows the centre of the composition, which is common to the tree approaches $\mathbb{R}^D_+$, $\mathcal{T}_p$, $\mathcal{T}_s$, with the centre of t_s and t_p. The inspection of Table 1 reveals that the total sum of the centre in $\mathbb{R}^D_+$ is not equal to the centre total in $\mathcal{T}_s$. However, t_p of the centre in $\mathbb{R}^D_+$ is equal to the centre t_p, as predicted by the isometry between $\mathbb{R}^D_+$ and $\mathcal{T}_p$. Moreover, the differences between the centres of samples B, A, BA appear to be substantial, indicating significant differences in the centres for all approaches. Table 1 also shows the metric variances for A, B, and BA samples in the three approaches. As expected, the approaches $\mathbb{R}^D_+$ and $\mathcal{T}_p$ give the same metric variances, which in turn differ from the metric variance in the $\mathcal{T}_s$ approach.

Table 1: Centres and metric variances of Phytoplankton data for samples B, A, and joint sample (BA). The upper table is the centre and metric variance in $\mathbb{R}^D_+$ in abundances (counts per liter); the sum and product total t_p of centre abundances are also shown. The lower table shows the compositional center, common to the three approaches $\mathbb{R}^D_+$, $\mathcal{T}_p$, and $\mathcal{T}_s$. The centre total in $\mathcal{T}_s$ is t_s, and t_p is the centre total in $\mathcal{T}_p$. The metric variances in the $\mathbb{R}^D_+$ approach are equal to those of $\mathcal{T}_p$.

	Mvar			**centre in $\mathbb{R}^D_+$**							
	$\mathbb{R}^D_+$, $\mathcal{T}_p$	t_s	t_p	ana	aph	ocy	aug	aud	act	ank	cry
B	7.268	4244	7573866	627	491	94	2373	67	193	287	111
A	6.867	7530	69151799	957	671	856	3597	108	320	715	305
BA	8.705	6079	32945104	830	604	408	3129	92	270	526	218
	Mvar			**centres in $\mathcal{T}$**							
	$\mathcal{T}_s$	t_s	t_p	ana	aph	ocy	aug	aud	act	ank	cry
B	6.039	5456	7573866	.148	.116	.022	.559	.016	.046	.068	.026
A	5.311	9654	69151799	.127	.089	.114	.478	.014	.043	.095	.041
BA	6.241	7973	32945104	.137	.099	.067	.515	.015	.045	.087	.036

The biplot representation in the three approaches provides a better insight about the features of the samples. Figure 2 (top) shows the biplot in the case $\mathbb{R}^D_+$; it is obtained from the singular value decomposition (svd) of the centered coordinates, i.e. the logarithms of abundances, which are plotted as variables. The middle panel of Figure 2 is the biplot for the $\mathcal{T}_p$ case. It has been obtained from the svd of the centered orthonormal coordinates in $\mathcal{T}_p$, i.e. the total t_p (labelled tprod) and the centered clr-coefficients of the composition. Comparison of the two biplots (top and middle) shows that the data points are exactly the same and also the projection in the two first principal components are equal, as predicted by the isometry of $\mathbb{R}^D_+$ and $\mathcal{T}_p$. On the other hand, the sample B (blue) and A (green) appear quite well separated in these two dimensional projections. Figure 2 (bottom) shows the biplot using $\mathcal{T}_s$. The centered logarithm of the sum of abundances and the centered clr's of the compositions are used as variables. The obtained projection is different from those obtained in Figure 2 (top and middle) and represents a lower percent of metric variance. In the $\mathcal{T}_p$ approach the first principal component is dominated by the total $\prod x_i^{1/\sqrt{D}}$ whereas the total sum is relatively less important in the first principal component in Figure 2 (bottom). This comparison is reinforced by the fact that the compositional part coincides in $\mathcal{T}_p$ and $\mathcal{T}_s$.

In the biplots in Figure 2 (middle) and (bottom) the totals appear almost orthogonal to the separation of samples B and A, i.e. they play an important role in the discrimination of the two samples, thus confirming that differences between B and A are not only compositional but also in total, whichever is this total. The traditional CoDa-biplot is enriched by the biplot in $\mathcal{T}_s$ as the total can be represented jointly with the standard clr variables.

MANOVA techniques can be used for testing equal centres in samples B and A. The vectors of abundances are represented in coordinates and then treated as multivariate real variables. Certainly, in the two approaches examined, $\mathbb{R}^D_+/\mathcal{T}_p$ and $\mathcal{T}_s$, the centres are clearly significantly different, as expected after the representation in the biplots. Also biplots in Figures 2 (middle) and 2 (bottom) suggest coordinates with significant and non-significant different means. For instance, the balance $\ln(\text{ana}/\text{aud})/\sqrt{2}$ has means in B and A which are not significative at all in an ANOVA F-test. Contrarily, balances between variables pointing at opposite sides of the axis ana-aud have significant different means. The log-totals also have significant different means for the two samples. As a conclusion, the hypothesis of equal centres for the samples B and A should be rejected with a very low significance (much less than 10^{-4}).

In the studied case, the relevant total was assumed to be the sum of the abundances, and the approach in $\mathcal{T}_s$ provides appropriate information for the study of the Phytoplankton data set. The approach in $\mathbb{R}^D_+$ also provides similar information, but the product total excessively dominates the analysis—as the first PC is essentially the product total.

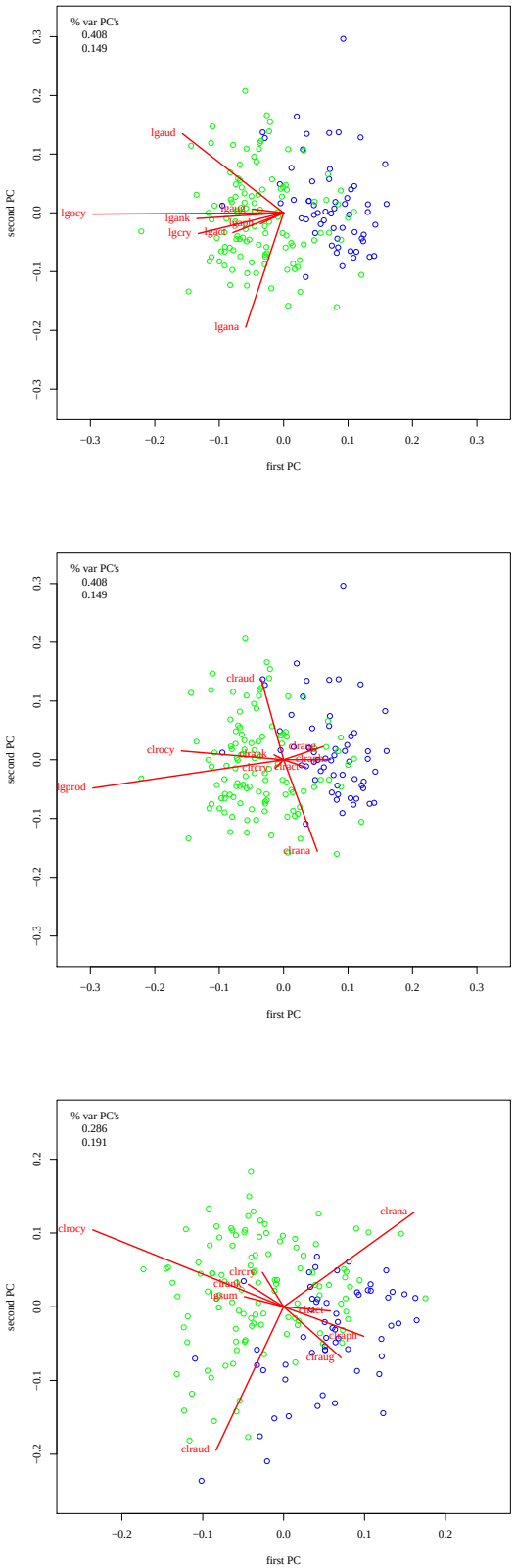


Figure 2: Phytoplankton data biplots. Sample B in blue; sample A in green. Top: $\mathbb{R}_+^D$ approach. Middle: $\mathcal{T}_p$ approach. Bottom: $\mathcal{T}_s$ approach.

ISBN: 978-3-200-03103-6

7 Conclusions

Data carrying compositional information plus a total are frequently analysed as vectors in $\mathbb{R}_+^D$ with the standard Euclidean structure, i.e. by taking logarithms and then using the standard multivariate methods designed for the real space $\mathbb{R}^D$. This practice is not compatible with the Aitchison geometry for the composition, obtained by closure, and the common idea that the sum of all components in $\mathbb{R}_+$ is a relevant total. Assuming the standard Euclidean structure of $\mathbb{R}_+^D$ (i.e. the log-geometry) and, simultaneously, that closed data are compositions, implies that the total is the product of all components to the power $1/\sqrt{D}$.

As a conclusion, if a data set in $\mathbb{R}_+^D$ is assumed compositional and the relevant total is the sum, it is advisable to perform the analysis in the product space $\mathcal{T} = \mathbb{R}_+ \times \mathcal{S}^D$. Sometimes, an analysis in $\mathbb{R}_+^D$ can give similar results to those obtained in $\mathcal{T}_s$, but circumstances in which they are similar are not clear.

Looking ahead, one of the central issues in compositional data analysis is subcompositional coherence (Aitchison, 1986) which raises the question as to how subcompositional coherence is reflected in $\mathbb{R}_+^D$ and $\mathcal{T}$. Also, while the basic approach to sampling zeros will be analogous to the usual approach in $\mathcal{S}^D$, a detailed study is needed to assess the impact of substitution techniques on the total function chosen for analysis.

Acknowledgements

This research has been supported by the Spanish Ministry of Education, Culture and Sports under a *Salvador de Madariaga* grant (Ref. PR2011-0290); by the Spanish Ministry of Economy and Competitiveness 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 project Ref: 2009SGR424.

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.

Fréchet, M. (1948). Les éléments aléatoires de nature quelconque dans une espace distancié. *Annales de L'Institut Henri Poincaré 10*(4), 215–308.

Lovell, D., W. Müller, J. Taylor, A. Zwart, and C. Helliwell (2011). Proportions, percentages, ppm: do the molecular biosciences treat compositional data right? In *Compositional Data Analysis: Theory and Applications.*, pp. 193–207. Pawlowsky-Glahn, V. and Buccianti A. (eds.), *Compositional Data Analysis: Theory and Applications*, Wiley, Chichester UK.

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.

Application of chemometric analysis to determine the degree of contamination in materials obtained by thermal conversion of biomass.

M. SAJDAK and S. STELMACH

Institute for Chemical Processing of Coal, Poland, msajdak@ichpw.zabrze.pl

1 Introduction

Biomass is the oldest and most widely used renewable source of energy. It is the third largest natural energy source (European Commission, 2000). Due to a continual increase in fossil fuel prices, renewable energy sources (RES), including biomass, are expected to become the second largest source of energy for electricity generation by the end of 2035. According to a report by the International Energetic Agency (IEA, 2012), global sources of bioenergy sufficiently secure biomass and biofuel supplies for energy demand, with minimal impact on food production. An increasing demand for biomass and products of biomass conversion (torrefied or carbonised biomass) may cause an unnatural increase in the calorific value of products, which are assumed to be natural and renewable energy carriers. Biomass may be doped with fossil fuels, polymers, and wastes from the furniture industry with significant contents of resins (polyesters, alkyds, and polyuretanes) or glues (urea glues and polyvinyl acetate glues), which are not classified as biomass. Analyses of carbonised batches of these solutions would prove time- and cost-consuming, considering that ^{14}C analysis is the best available technique for determining the origin of a sample. To fulfill current market requirements, a new combination of chemometric methods and conventional analyses of solid fuels (elementary analysis, heat of combustion, and oxide content in ashes) are suggested to simply and rapidly evaluate the purity of thermal conversion products (Sajdak, 2012, Sajdak et al., 2012). Certain methods have already been applied in the classification of various types of biomass (Guangcan Tao et al., 2012a, b). However, the use of this technique for purity control of biomass and products of biomass thermal conversion is challenging. The research presented in this paper applied a combination of principal component analysis (PCA), classification and regression trees (C&RT), and hierarchical clustering analysis (HCA). HCA enables the quantitative estimation of investigated relations, which facilitates the identification of pollutants in biomass and products of biomass thermal conversion.

PCA was employed to determine the variables necessary for a clear description of the types of tested materials and to present the interactions between these variables and their groups. The C&RT method facilitated the creation of an algorithm for testing the material classifications. Regression analysis is applied to estimate the pollution content in samples with compositions that are significantly different from natural samples. The presented methodology requires continuous improvement to enhance the optimisation of the applied techniques.

2 Experimental procedure

2.1 Materials and thermal conversion method

Granules of wood biomass (pine) and polypropylene without any filler, with a radius of 4 mm and coloured with charcoal (Daplen provided by Borealis), were utilised in this research. The tested biomass was pre-dried at 80°C to attain 2–3% water content (W_{tr}) and grinded to attain a granulation of <5 mm. The pure samples of biomass and various mixtures of biomass with PP (10% to 58% of PP) were subjected to a pyrolysis process. The prepared samples were pyrolysed under various thermal conditions, which are established in the central composite design by the design of experiment method (DOE, Table 1).

In each test, 150 g of the sample was placed in a steel retort and pyrolysed in a nitrogen atmosphere. Figure 1 shows a schematic of the biomass thermal conversion work station. Nitrogen was purged

through the bed from the bottom of the retort with a constant flow of 3 dm^3/h. Prior to heating, the sample was washed with nitrogen for 15 min to attain a neutral atmosphere in the retort. The retort was subsequently heated at the rate of 5°C/min. Once the final temperature was obtained, the sample was maintained under these conditions for 50 min and then cooled to room temperature. The basic parameters of planned and performed experiments are listed in table 1.

Table 1 Matrix of experiments for the pyrolysis of biomass and biomass-PP blends.

	No.	Temp. [°C]	PP [%]
Regular experiment	1	500	10
	2	500	50
	3	700	10
	4	700	50
	9	458	30
	6	740	30
	7	600	0
	8	600	58
Central	9	600	30
	10	600	30
	11	600	30
Blank	12	500	0
	13	600	0
	14	700	0
	15	458	0

Similar research exists that focuses on biomass pyrolysis under various thermal conditions (Słowik K. et al. 2011). However, this study considered only pure biomass without the addition of other materials. All samples of carbonised biomass were investigated using the FT-IR analysis and the DRIFT method.

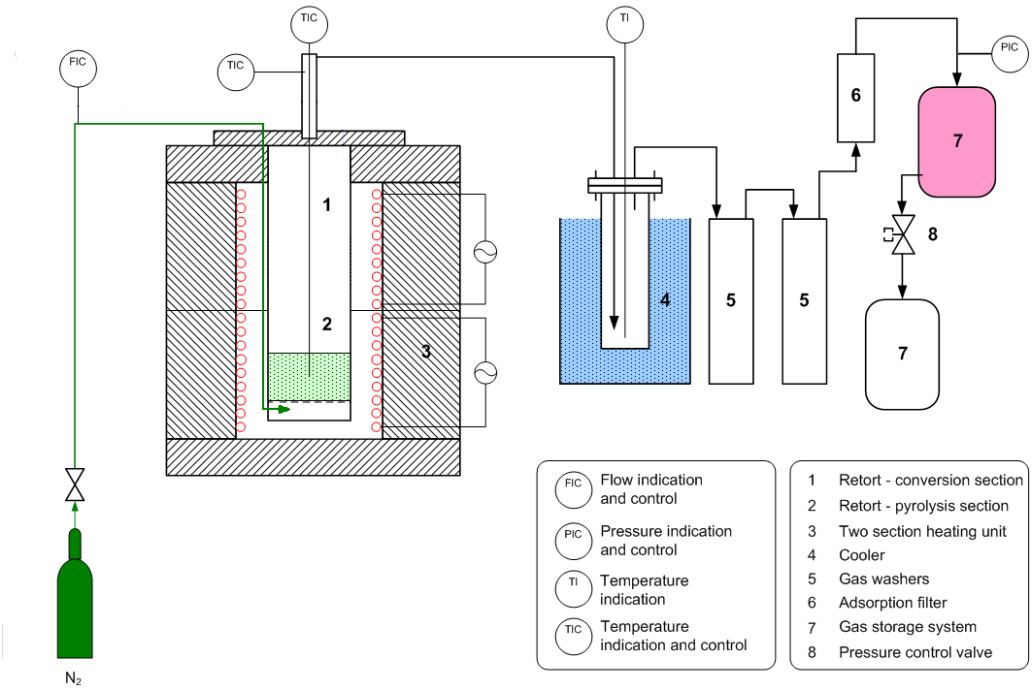


Figure 1 Diagram of the experimental stand for biomass thermal conversion testing

2.2 Analytical method used to study bio-char

Infrared spectra were examined using the DRIFT technique (Diffuse Reflectance Infrared Fourier Transform) with a Tensor 27 spectrometer provided by Bruker. DRIFT spectra measure the vibrations of functional groups on a sample surface. A test tablet was prepared with 300 mg of potassium bromide (KBr) and 2–3 mg of the test sample; both were weighed on analytical balances. The KBr was grinded in a mortar and placed in a metal pot (capsule), which was placed in a column in an adapter. The KBr surface was smoothed with a spatula. The background noise was measured in 4 cm^{-1} over 128 scans. The next stage involved the preparation of a mixture of the test sample with KBr. The sample of carbonised biomass was grinded with KBr and placed in a metal pot according to the abovementioned procedure. Measurements were performed under the same parameter conditions as the conditions for the background noise. Data were extracted from 3 measurements by twisting the sample by a small angle each time. The results were averaged. Figure 2 presents the exemplary spectrograms of the analysed carbonised biomass with PP at 500°C (0% PP, 10% PP, and 50% PP).

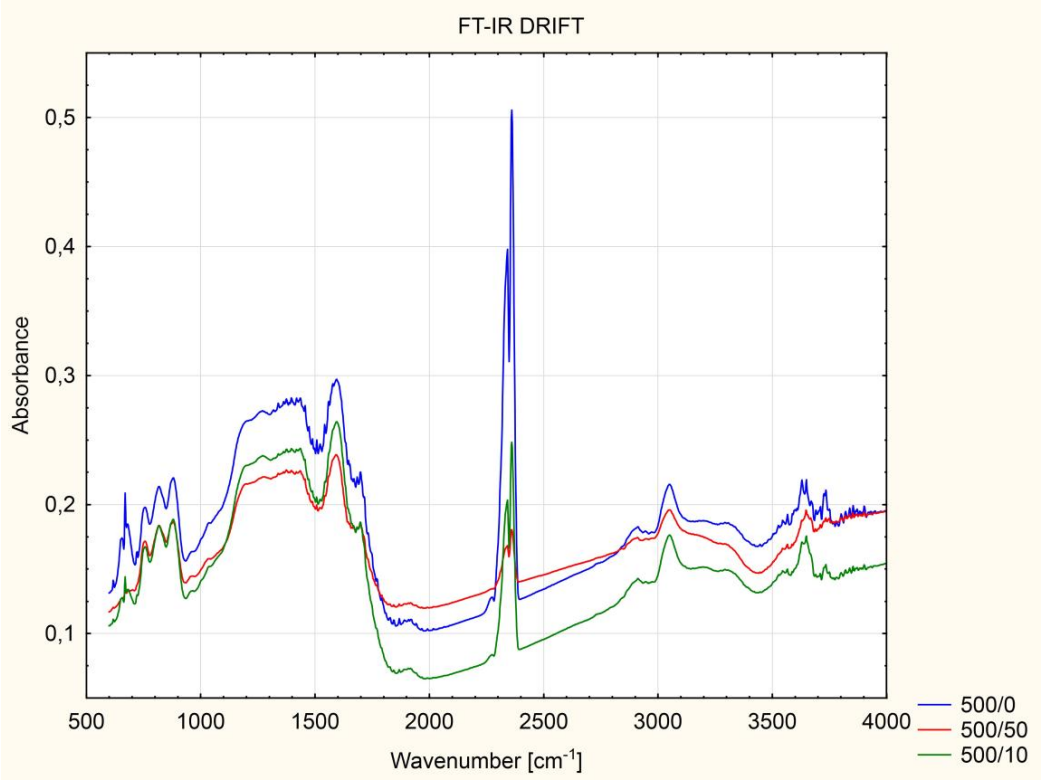


Figure 2 Example spectrogram of the analysed char from the biomass and polypropylene

2.3 Chemometric and compositional data analysis used to study bio-char

Prior to the chemometric analysis, FT-IR spectral data were normalised according to equation (1). This procedure was essential for additional chemometric analyses and for proper interpretation of the results (Nisbet R., et al. 2009). This approach aims to unify the variables and variance in the test data, which are set to values between 0 and 1. Equation (1) is expressed as

$$x_{ij} = \frac{a_{ij} - b_j}{s_j} \qquad (1)$$

where

x_{ij} standardised parameter value,

a_{ij} initial value of the parameter,
b_j average value of the parameter, and
s_j standard deviation of the jth parameter.

Figure 3 presents the results of the spectral data normalisation process that was performed according to equation (1). After normalisation of the spectral data of the carbonised samples (figure 3a), a new data set was obtained (figure 3b) that can be subsequently normalised and analysed. The normalised data set was subjected to cluster analysis and principal component analysis.

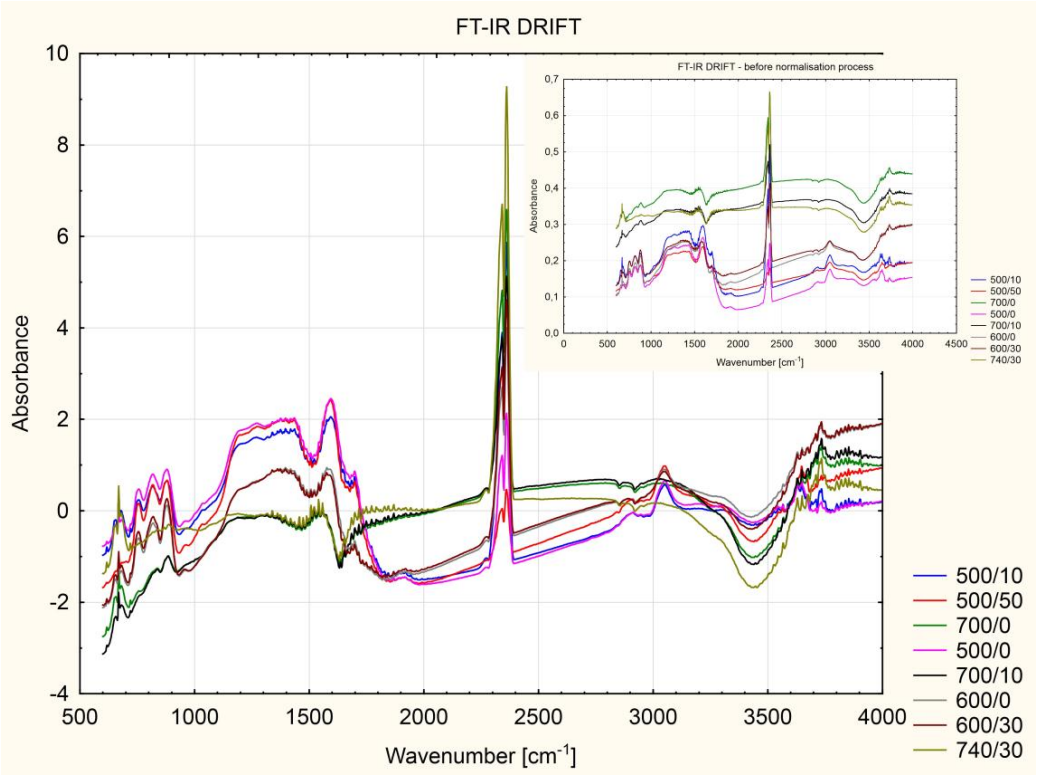


Figure 3 Spectra of the investigated materials (bio-char) before (a) and after (b) the normalisation process

After pre-preparation, the data set was subjected to cluster analysis (CA) and principal component analysis . Cluster analysis is an exploratory multivariate method that can be used to describe the relationships among variables. For the hierarchical clustering analysis (HCA), Ward's method was employed to obtain the cluster plots. This method, which is considered extremely efficient, utilises an analysis-of-variance approach to evaluate the distance between clusters and minimises the sum of squares (SS) among clusters. The Euclidean distance, which was chosen for the analysis in this study, is expressed as

$$d_{ij} = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk}\right)^2} \qquad (2)$$

where
m number of variables
d_{ij} Euclidean distance between objects i and j,
x_{ik} value of the variable i,
x_{jk} value of the variable j

Ward's method differs from other binding methods in the use of a variance analysis to estimate the distance between clusters. This method minimises the deviation in the sum of squares (SS) between arbitrary clusters.

Direct analysis of large data sets (e.g., spectral data) is challenging. Therefore, a simpler interpretation of the data from the cluster analysis is presented in dendrograms rather than in tables, as shown in figure 4.

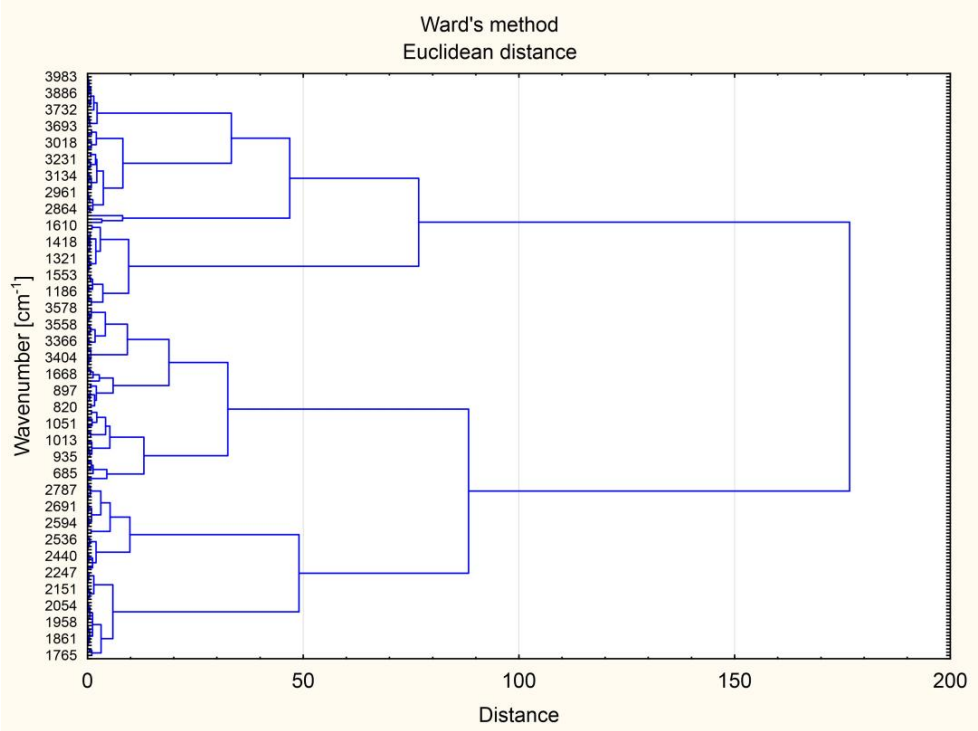


Figure 4 Horizontal hierarchical tree plot.

Principal component analysis is another technique employed in this research. It is an efficient tool for the analysis of multidimensional data because it defines the support vectors from the covariance matrix of original variables. Facilitates the discovery of relationships between the test variables and the materials (samples) and reduces the matrix dimension (reduction of variable number is significant in the subsequent analysis). A characteristic feature of PCA is the ability to determine dominant components (specific value) in decreasing order from information conveyed by subsequent variables. In some cases, a new variable may become a specific physical or chemical interpretation during PCA (Tao at al., 2012 a, b).

Principal component analysis is primarily applied for the following reasons:

- Data space reduction (fewer variables),
- Transformation of input variables into new variables—principal components, and
- Graphical presentation of multidimensional data on a 3D surface without significant loss of information.

Data obtained from principal component analysis were applied as input values in the mathematical model evaluation, which was based on classification and regression trees and constructed to estimate the polymer content in the biomass pyrolysis.

The primary objective of the C&RT method is a gradual division of test arguments (samples) into homogeneous subsets, until homogeneity is attained based on assignment to the correct class. A set of rules based on the "if" formula is created, which enables the subdivision of arguments into groups, depending on subsequent qualitative and quantitative parameters. In the C&RT method, a decisive formula may be presented graphically in the form of a decisive tree. The tree resembles a graph that is comprised of a mother root from which a minimum of two paths to nodes emerges.

The tree resembles a graph that consists of a root node and a minimum of two branches, which emerge from the root node and lead to inferior nodes (child nodes). Each node corresponds to a class

description, and each branch corresponds to a decision rule, i.e., a condition that is related to arguments from an entry data set and describes the case in which each branch is chosen. Child nodes become parent nodes during successive splitting of the data set. Each division is performed for separate features (parameters). In a common algorithm, the conditions on the branches of each node must be complementary, such that one possible path is provided downward when 'walking the tree'. Nodes that do not contain any child nodes are known as leaf nodes, outer nodes or terminal nodes; they represent the final classes.
In a formal approach, a tree is a graph without cycles (loops), which displays only one path between two different nodes.

The main advantages of C&RT are as follows:

- Ability to present arbitrarily complex problems.
- Resistance to unusual feature values.
- Resistance to a large number of features that do not influence the outgoing variable.

The main disadvantages of C&RT are as follows:

- Large size of the trees, because the algorithm tests only one feature at a time.

Each node corresponds to a proper class and each class is defined by a function (decisive parameter). In subsequent division, a child node becomes a parent node, which undergoes splitting.

3 Results and discussion

FT-IR spectral data of the examined carbonised biomass were subjected to cluster analysis and principal component analysis after the initial normalisation. Each chemometric and statistic operation described in this paper was conducted using STATISTICA software (StatSoft, 2012).

After cluster analysis, three data subsets were established and characterised by different temperatures. In this analysis, the results are presented in the form of a tree diagram—a dendrogram (figure 5)—instead of in a common tabularised form. The first group (figure 5, green) presents samples obtained from pyrolysis at a maximum temperature of 500°C. The second group (figure 5, red) describes the samples obtained from pyrolysis at an approximate temperature of 600°C. Significant differences in the distances between the tested samples can be attributed to the different contents of PP in the mixture with biomass. Unstable conditions during pyrolysis were unavoidable (figure 5, red).

The last, and third, group (figure 5, red) depicts the samples obtained from pyrolysis at a temperature above 700°C, with one exception: a sample of biomass with 58% PP content pyrolysed at 600°C. This excessive quantity of polymer changes the process conditions, which creates a more processed material.

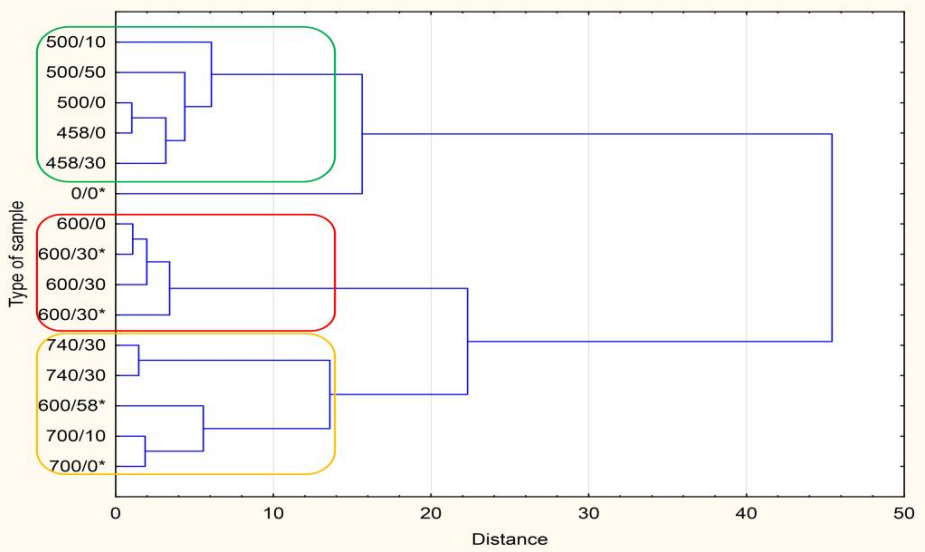


Figure 5 Horizontal hierarchical tree plot.

The polymer applied in the subsequently described research (Borealis Daplen) undergoes depolimerisation (thermal decomposition) at approximately 400°C. Elemental analysis was performed by direct method using Macro Cube CHNS, O, and a Cl elemental analyser provided by Elementar Analsensysteme GmbH (high-temperature pyrolysis), and oxygen detection was performed by the ND-IR method (Muzyka R., et al., 2013).

When the cluster analysis was completed, the obtained data were subjected to chemometric analysis—principal component analysis. As a result, a new data set with fewer variables was established; the tree's basic variables constituted the main components. The first (PC 1) and third (PC 3) components allow descriptions of the given data set with respect to the pyrolysis conditions in which the samples were processed. The second component (PC 2) refers to the PP content in the co-pyrolysis with biomass (figure 6). The new data set from the PCA was used as the input data in the analysis based on classification and regression trees.

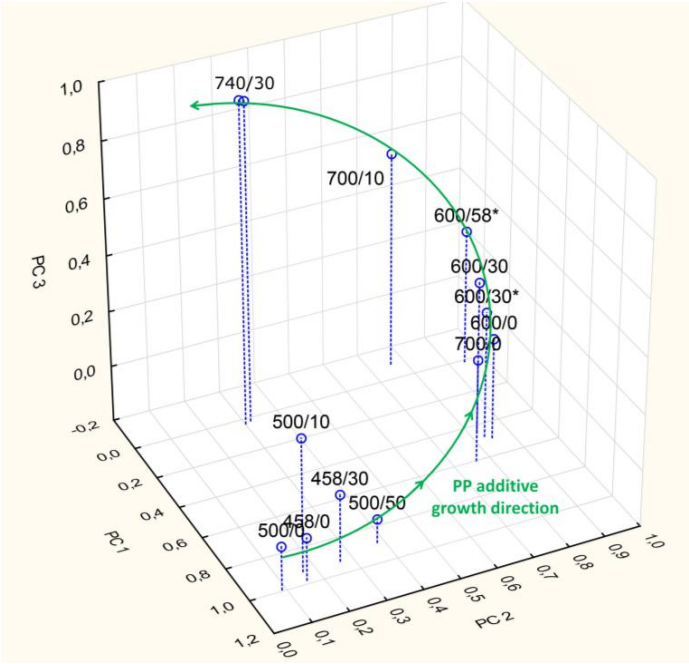


Figure 6 The projection of samples on the space defined by the tree's principal factors

The data set obtained using the main component method during the analysis of examined materials are presented in table 2

Table 2 Results of principal component analysis.

No	PC 0	PC 1	PC 2	PC 3	Temp. [°C]	PP [%]
1	0,58	0,79	-0,06	-0,11	500,00	10,00
2	0,53	0,75	0,16	0,27	500,00	50,00
3	0,41	0,90	-0,07	0,16	500,00	0,00
4	0,72	-0,29	0,44	-0,34	700,00	10,00
5	0,82	0,28	0,49	0,09	600,00	0,00
6	0,80	0,39	0,42	0,08	700,00	0,00
7	0,84	0,25	0,46	0,01	600,00	30,00
8	0,63	-0,13	0,04	-0,77	740,00	30,00
9	0,76	-0,20	0,60	-0,04	600,00	58,00
10	0,86	0,21	0,45	-0,08	600,00	30,00
11	0,54	0,80	0,05	0,12	458,00	30,00
12	0,63	-0,14	0,06	-0,75	740,00	30,00
13	0,45	0,87	-0,02	0,17	458,00	0,00

The new data set (table 2) from the PCA was employed as input data in the analysis based on classification and regression trees. Using this method, the predictive C&RT model was constructed to estimate the PP content added to pyrolysed biomass, as presented in figure 7.

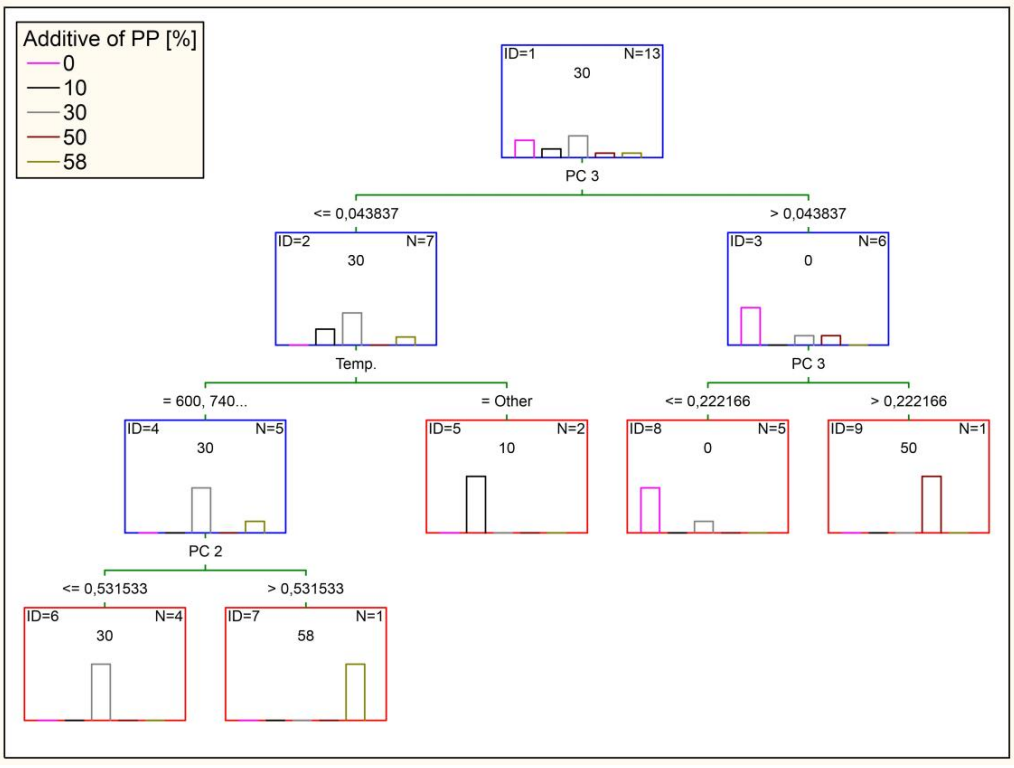


Figure 7 The preliminary predictive model based on the C&RT method

Although it can be used to estimate the PP content in a sample of carbonised biomass, the presented model is a preliminary model due to the small amount of input data employed during its

creation. Therefore, the model was not subjected to cross-check validation and is considered a model suitable for the initiation of new research.

As a result of this research, a procedure pattern was formulated for these types of analyses. The procedure algorithm is shown in figure 8.

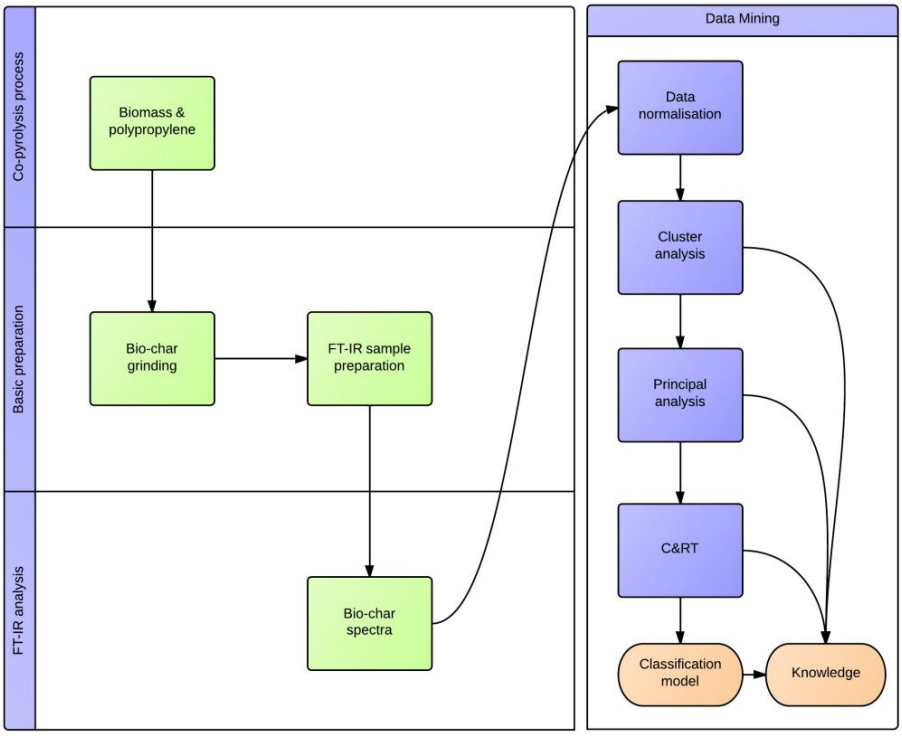


Figure 8 Algorithm proceedings with data

4 Conclusion

This paper presents the results of research on the application of chemometric methods, which can be employed to determine the presence and quantity of polymeric materials carbonised with biomass in pyrolytic processes. The cluster analysis employed in the research allowed the creation if three well-defined groups of biomass, which were carbonised at temperatures of 500°C, 600°C and 750°C. The results exhibit 100% agreement with the initial assumptions. The Euclidean distance was used to describe the similarities between the tested materials, which demonstrated the potential of the new method as a quantitative analysis of the PP content in carbonised biomass.

The research results indicate the potential of its application to construct a complete analytical procedure. In the future, this procedure will facilitate the identification of pollutions in carbonised biomass that is available on the market.

The presented data are preliminary data. However, based on this research, a direction for subsequent analyses in this field has been defined.

5 Acknowledgments

These studies were funded under the budget of research task No. 4 "Development of integrated technology of fuels and energy from biomass, agricultural waste, and others as part of the strategic program of research and development: Advanced technology of obtaining energy" provided by the NCBiR.

6 References

Muzyka R., Topolnicka T., „Bezpośrednie oznaczanie zawartości tlenu w węglach z wykorzystaniem metody wysokotemperaturowej pirolizy", Przegląd Górniczy, 3/2013 (In Press, In Polish)

Nisbet R., Elder J.F., Miner G., (2009) Handbook of statistical analysis and data mining applications (Academic Press, Amsterdam)

Sajdak M. (2012) „Application of chemometrics to identifying solid fuels and their origin", Central European Journal of Chemistry Volume 11, Number 2 (2013), 151-159, DOI: 10.2478/s11532-012-0145-8.

Sajdak M. and Piotrowski O. (2012), „C&RT model application in classification of biomass for energy production and environmental protection", Central European Journal of Chemistry Volume 11, Number 2 (2013), 259-270, DOI: 10.2478/s11532-012-0147-6.

Sharypov, V.I., N.G. Beregovtsova, B.N. Kuznetsov, L. Membrado, V.L. Cebolla, N. Marin and J.V. Weber, (2003). Co-pyrolysis of wood biomass and synthetic polymers mixtures. Part 111: Characterization of heavy products. J. Anal. Applied Pyrolysis, 67: 325-340.

Słowik K., Stelmach S. (2011) „Wpływ warunków procesowych na uzysk produktów pirolizy słomy rzepakowej", Innowacyjne i przyjazne dla środowiska techniki i technologie przeróbki surowców mineralnych 225-234 (In Polish)

StatSoft, Inc., USA, (2009), STATISTICA data analysis software system, version 10

Tao G., at al. (2012a), Biomass properties in association with plant species and assortments I: A synthesis based on literature data of energy properties, Renewable and Sustainable Energy Reviews, Volume 16, Issue 5, June 2012, Pages 3481-3506, ISSN 1364-0321, 10.1016/j.rser.2012.02.039.

Tao G., at al. (2012b), Biomass properties in association with plant species and assortments. II: A synthesis based on literature data for ash elements, Renewable and Sustainable Energy Reviews, Volume 16, Issue 5, June 2012, Pages 3507-3522, ISSN 1364-0321, 10.1016/j.rser.2012.01.023.

World Energy Outlook 2012; IEA France 2012.

Regression between compositional data sets

R. TOLOSANA-DELGADO[1] and K.G. VAN DEN BOOGAART[1,2]

[1]Department of Modelling and Valuation - Helmholtz Institute Freiberg for Resources Technology, Germany
[2]Institute for Stochastics - Technical University Bergakademie Freiberg, Germany
r.tolosana@hzdr.de

Abstract

Linear regression where both the explained and the explanatory variables form compositions are naturally tractable within the log-ratio framework. Fitting such models does not imply any difficulty: they can be fit in a standard way after applying any one-to-one logratio transformation to each compositional set. Problems arise to test and display the model, due to the large dimension of the model parameters space, and the difficult interpretation of classical hypotheses in terms of the original components. This contribution proposes two graphical representations of the model: in the form of a biplot, parallel to redundancy analysis, and as confidence ellipses on the parameters projected onto a set of subcompositions. Each of these representations brings also associated a way to test for certain subcompositional independence hypotheses. An exact, general, Scheffé-like test of independence (for the whole composition or any subcomposition) can be derived from a generalized eigenvalue problem of the matrix of regression coefficients and its estimation covariance matrix. For certain hypotheses of independence, classical tests based on Hotelling's T^2 or χ^2 distributions can also be adapted. Any of these tests can be used to calculate the radii of confidence ellipses on the parameters, in order to visualize the corresponding tests. This provides a toolbox to reduce the complexity of compositional-to-compositional regression, and enables a structured way of exploring and testing which components of the explanatory set influence which components of the explained set.

1 Introduction

Linear regression is one of the grounding techniques of statistics, aimed at detecting and quantifying the dependence between two sets of variables, one considered explanatory and the other considered as dependent or explained. This contribution focuses on the case where both the dependent and independent sets are compositions. The main problem of this setting is the complexity of the resulting regression system, where all explanatory components are involved in the prediction of any aspect of the explained composition. This often implies estimating (too) many parameters. Both for prediction and explanation uses, it would be better to restrict the dependence to a subset of components. This paper presents a set of graphical tools and statistical tests devised to adequately explore this subcompositional independence. This is illustrated with two examples, one from mining, and one linking educational level to political preferences.

2 Notation and grounding geometric concepts

Let $\mathbf{X} = [x_{ni}]$ and $\mathbf{Y} = [y_{nj}]$ be two data sets, with $n = 1, \ldots, N$ paired observations and respectively $i = 1, \ldots, P$ and $j = 1, \ldots, Q$ components. The rows of each matrix $\mathbf{x}_n = [x_{n1}, x_{n2}, \ldots, x_{nP}]$ and $\mathbf{y}_n = [y_{n1}, y_{n2}, \ldots, y_{nQ}]$ are compositions of different simplexes, $\mathbf{x}_n \in \mathbb{S}^P$ and $\mathbf{y}_n \in \mathbb{S}^Q$, such that the pair $(\mathbf{x}_n, \mathbf{y}_n) \in \mathbb{S}^P \times \mathbb{S}^Q$ belongs to the cartesian product of the two simplexes. For each of these spaces, operations of perturbation $\oplus$, powering $\odot$ and the Aitchison scalar product $(\cdot,\cdot)_a$ can be defined, which induce Euclidean space structures to each of $\mathbb{S}^P$, $\mathbb{S}^Q$ and $\mathbb{S}^P \times \mathbb{S}^Q$ (Aitchison, 1986; Pawlowsky-Glahn and Egozcue, 2001; Aitchison, 2002). What follows is valid for any of these three spaces.

Perturbation (Aitchison, 1986) and powering (Aitchison, 2002) are respectively the closed component-wise product of two compositions $\mathbf{a}, \mathbf{b} \in \mathbb{S}^D$ of the same space, and the closed component-wise powering of a composition by a scalar:

$$\mathbf{a} \oplus \mathbf{b} = \mathcal{C}[a_1 b_1, a_2 b_2, \ldots, a_D b_D], \quad \lambda \odot \mathbf{a} = \mathcal{C}[a_1^\lambda, a_2^\lambda, \ldots, a_D^\lambda].$$

In both cases, $\mathcal{C}[\cdot]$ represents the closure operation, which divides its argument vector by its total sum and forces the constant sum constraint to 1,

$$\mathcal{C}[\mathbf{a}] = \frac{1}{\mathbf{a} \cdot \mathbf{1}^t}\mathbf{a},$$

where $\mathbf{1}$ is a row-vector of ones. Actually, in this contribution all vectors are considered row-vectors, for consistency with the literature on compositional data analysis. The Aitchison (2002) scalar product $(\cdot,\cdot)_a$ is proportional to the scalar product of all possible pairwise logratios,

$$(\mathbf{a},\mathbf{b})_a = \frac{1}{2D}\sum_{i>j}^{D} \ln\frac{a_i}{a_j}\ln\frac{b_i}{b_j}.$$

As is well-known, compositions convey only relative information, which can be extracted with any set of one-to-one logratios of its components. The first option is the centered log-ratio transformation (Aitchison, 1986),

$$\mathrm{clr}(\mathbf{b}) = \ln(\mathbf{b}) - \frac{1}{D}(\ln(\mathbf{b})\cdot\mathbf{1}^t)\cdot\mathbf{1},$$

i.e. equal to the component-wise logarithm of the composition, up to a constant value that corresponds to the mean of th D transformed components. For this fact, the clr coefficients sum up to zero, and they can be identified with points on a hyperplane $\mathbb{H}\in\mathbb{R}^D$ of dimension $D-1$ orthogonal to the vector $\mathbf{1}$. The clr transformation is an isometry between $\mathbb{S}^D$ and $\mathbb{H}$ (Egozcue et al., 2003), with inverse

$$\mathrm{clr}^{-1}(\boldsymbol{\beta}) = \mathcal{C}[\exp(\boldsymbol{\beta})].$$

The clr representation of a composition is often convenient because each transformed coefficient can be related to an original component. However, other representations may be fitter for a certain purpose. For instance, the pairwise logratio of components i and j can be obtained by

$$\beta_{ij} = \mathrm{clr}(\mathbf{b})\cdot\mathbf{w}_{ij}^t = \ln(\mathbf{b})\cdot\mathbf{w}_{ij}^t,$$

where $\mathbf{w}_{ij}$ is a vector of zeros, except a $+1$ in the i-th position and a -1 in the j-th one. Actually, any direction $\mathbf{w}$ of $\mathbb{H}$ can be used to project the composition, and the resulting score may be calculated as

$$\beta_{\mathbf{w}} = \mathrm{clr}^t(\mathbf{b})\cdot\mathbf{w}^t = \sum_{i=1}^{D} w_i\cdot\ln(b_i) = \ln\prod_{i=1}^{D} b_i^{w_i}.$$

Note that a vector $\mathbf{w}$ represents a direction within $\mathbb{H}$ if it is orthogonal to $\mathbf{1}$, i.e. if its components add to zero. Finally, one can choose a set of $D-1$ vectors forming a basis of $\mathbb{H}$ and use the coordinates of the clr-transformed composition with respect to it. These can be computed as

$$\boldsymbol{\beta}_{\mathbf{W}} = \mathrm{clr}(\mathbf{b})\cdot\mathbf{W}^t = \ln(\mathbf{b})\cdot\mathbf{W}^t$$

where each row $\mathbf{w}_i$ of $\mathbf{W}$ is a D-component vector of $\mathbb{H}$. If the basis is orthogonal, i.e. if $\mathbf{W}\cdot\mathbf{W}^t=\mathbf{I}$, then the resulting transformation is called an *isometric log-ratio transformation* (ilr Egozcue et al., 2003), and as a one-to-one representation of compositional data, it can be written as:

$$\mathrm{ilr}(\mathbf{b}) = \ln(\mathbf{b})\cdot\mathbf{W}^t = \boldsymbol{\beta};\quad \mathrm{ilr}^{-1}(\boldsymbol{\beta}) = \mathcal{C}(\exp(\boldsymbol{\beta}\cdot\mathbf{W})) = \mathbf{b}.$$

This establishes also a correspondence between ilr coordinates and clr coefficients,

$$\mathrm{ilr}(\mathbf{b}) = \mathrm{clr}(\mathbf{b})\cdot\mathbf{W}^t;\quad \mathrm{clr}(\mathbf{b}) = \mathrm{ilr}(\mathbf{b})\cdot\mathbf{W}. \tag{1}$$

To simplify notation, we do not specify the basis used in the notation of ilr coordinates, i.e. $\boldsymbol{\beta}_{\mathbf{W}}=\boldsymbol{\beta}$, if this can be implied from the context. On the other hand, as we are working with two different simplexes, we have two different families of ilr transformations: those acting on the rows of $\mathbf{X}$ and those acting on the rows of $\mathbf{Y}$. To distinguish between them, the transformations may be denoted as $\mathrm{ilr}_x(\cdot)$ and $\mathrm{ilr}_y(\cdot)$ respectively, and their associated basis matrices as $\mathbf{W}_x$ and $\mathbf{W}_y$.

3 The regression model

3.1 Estimation in coordinates

An affine linear function is established to predict the rows of $\mathbf{Y}$ from those of $\mathbf{X}$,

$$\hat{\mathbf{y}}_n = \mathbf{b}_0 \oplus \mathrm{B}\mathbf{x}_n,\quad \mathbf{y}_n = \hat{\mathbf{y}}_n \oplus \boldsymbol{\epsilon}_n,$$

where B is a linear application from $\mathbb{S}^P$ to $\mathbb{S}^Q$, and the intercept $\mathbf{b}_0$, each prediction $\hat{\mathbf{y}}_n$ and each residual $\boldsymbol{\varepsilon}_n$ belongs to $\mathbb{S}^Q$ the simplex of $\mathbf{y}_n$. It is required to estimate $\mathbf{b}_0$, B and the spread of $\boldsymbol{\varepsilon}_n$.

On choosing a basis of each simplex, this regression model becomes

$$\mathrm{ilr}_y(\hat{\mathbf{y}}_n) = \boldsymbol{\beta}_0 + \mathrm{ilr}_x(\mathbf{x}_n)\cdot\mathbf{B}, \quad \mathrm{ilr}_y(\mathbf{y}_n) = \mathrm{ilr}_y(\hat{\mathbf{y}}_n) + \mathrm{ilr}_y(\boldsymbol{\varepsilon}_n), \tag{2}$$

where $\boldsymbol{\varepsilon}_n = \mathrm{ilr}_y(\boldsymbol{\epsilon}_n)$, $\boldsymbol{\beta}_0 = \mathrm{ilr}_y(\mathbf{b}_0)$ and $\mathbf{B}$ is the matrix representation of the application B in the two bases chosen.

Assuming that the ilr-transformed data are jointly normally distributed, classical multivariate regression methods provide estimates of each coefficient in $\boldsymbol{\beta}_0$ and $\mathbf{B}$, as well as its variance-covariance matrix $\boldsymbol{\Sigma}_{\mathbf{B}}$,

$$\hat{\mathbf{B}} = \left(\widehat{\mathrm{var}}\left[\mathrm{ilr}(\mathbf{X})\right]\right)^{-1}\cdot\widehat{\mathrm{cov}}\left[\mathrm{ilr}(\mathbf{X}),\mathrm{ilr}(\mathbf{Y})\right], \tag{3}$$

$$\hat{\boldsymbol{\beta}}_0 = \overline{\mathrm{ilr}(\mathbf{Y})} - \overline{\mathrm{ilr}(\mathbf{X})}\cdot\hat{\mathbf{B}}, \tag{4}$$

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{B}} = \frac{1}{N-1}\left(\widehat{\mathrm{var}}\left[\mathrm{ilr}(\mathbf{X})\right]\right)^{-1}\otimes\hat{\boldsymbol{\Sigma}}_{\epsilon} \quad (\mathbf{B}\text{ stacked by columns}), \tag{5}$$

where $\otimes$ denoted the Kronecker product, and $\hat{\boldsymbol{\Sigma}}_{\epsilon}$ is the variance of the ilr-residuals, estimated as $\hat{\boldsymbol{\Sigma}}_{\epsilon} = (N-P)^{-1}\cdot(\mathrm{ilr}(\hat{\mathbf{Y}}) - \mathrm{ilr}(\mathbf{Y}))^t\cdot(\mathrm{ilr}(\hat{\mathbf{Y}}) - \mathrm{ilr}(\mathbf{Y}))$. The estimation covariance matrix $\hat{\boldsymbol{\Sigma}}_{\mathbf{B}}$ allows to compute confidence areas for the parameters, and predictive areas for model predictions, both expressed in ilr coordinates. Note that in these expressions estimated variances and covariances of both $\mathrm{ilr}(\mathbf{X})$ and $\mathrm{ilr}(\mathbf{Y})$ are assumed to be the unbiased versions, i.e. divided by $N-1$.

3.2 Compositional interpretation

The intercept $\boldsymbol{\beta}_0$ and the residuals can be expressed back as compositions, simply by

$$\hat{\mathbf{b}}_0 = \mathrm{ilr}_y^{-1}(\hat{\boldsymbol{\beta}}_0),$$
$$\hat{\boldsymbol{\epsilon}}_n = \mathrm{ilr}_y^{-1}(\hat{\boldsymbol{\varepsilon}}).$$

For the model coefficients and its estimation variance expressions are a bit more complex. In Eq. (2), the rows of $\mathbf{B}$ identify directions of $\mathbb{S}^Q$, and its columns can be related to directions of $\mathbb{S}^P$. These may be better expressed in terms of a clr representation,

$$\mathbf{B}^c = \mathbf{W}_x^t\cdot\mathbf{B}\cdot\mathbf{W}_y.$$

Its estimation covariance matrix can also be expressed in clr coefficients, as

$$\hat{\boldsymbol{\Sigma}}_{\mathbf{B}}^c = (\mathbf{W}_x^t\otimes\mathbf{W}_y^t)\cdot\hat{\boldsymbol{\Sigma}}_{\mathbf{B}}\cdot(\mathbf{W}_x\otimes\mathbf{W}_y) \quad (\mathbf{B}\text{ stacked by columns}),$$

thanks to the distributivity and inversion properties of the Kronecker product. In the same way, It can be shown that all these expressions give the same results in clr or compositions, regardless of the actual ilr transformation used to obtain the estimates. This implies that B (and its estimation variance matrix) is an object with an intrinsic meaning, as $\mathbf{B}^c$ or $\mathbf{B}$ are only different representations of it.

4 Redundancy analysis

4.1 Singular directions of the matrix of coefficients

Given that B is a linear application from $\mathbb{S}^P$ onto $\mathbb{S}^Q$, the spectral theorem (Eaton, 1983) ensures the existence of a pair of bases $(\mathbf{V}_x,\mathbf{V}_y)$ on $\mathbb{S}^P$ and $\mathbb{S}^Q$ respectively, such that its matrix expression $\mathbf{B}$ is a part of a diagonal matrix, i.e. $b_{ij} = d_i\delta_{ij},$, with $i = 1,\ldots,P-1; j = 1,\ldots,Q-1$. These can be found with the singular value decomposition (SVD) of any of its matrix representations, e.g.

$$\mathbf{B}^c = \mathbf{V}_x^t\cdot\mathbf{D}\cdot\mathbf{V}_y, \tag{6}$$

where the rows $\mathbf{v}_{x,i}$ of $\mathbf{V}_x$ and $\mathbf{v}_{y,i}$ of $\mathbf{V}_y$ are called respectively left and right singular vectors of B, and the matrix $\mathbf{D} = [d_i\delta_{ij}]$ contains the singular values in the diagonal and is the representation of B in that sought basis. By plugging Eqs. (4) and (6) into the model (2) and using the ilr's defined by the SVD, one obtains

$$\mathrm{clr}(\hat{\mathbf{y}}_n)\cdot\mathbf{v}_{y,i}^t = \mathrm{E}\left[\mathrm{clr}(\mathbf{Y})\right]\cdot\mathbf{v}_{y,i}^t + d_i\cdot(\mathrm{clr}(\mathbf{x}_n) - \mathrm{E}\left[\mathrm{clr}(\mathbf{X})\right])\cdot\mathbf{v}_{x,i}^t, \quad i = 1,2,\ldots,R,$$

with $R \leq \min(P,Q) - 1$ the rank of B, i.e. the dimension of its image subspace, or the number of its non-zero singular values. Thus, expressed in these bases, a unit length of the projection of the departure of $\mathbf{x}_n$ from the mean of $\mathbf{X}$ onto the direction of $\mathbf{v}_{x,i}$ becomes d_i units of departure of $\mathbf{y}_n$ from the mean of $\mathbf{Y}$ along the direction of $\mathbf{v}_{y,i}$.

4.2 A biplot representation

This interpretation allows to represent the linear application B in the form of a biplot (Eckart and Young, 1936; Gabriel, 1971), i.e. a 2-rank approximation $\mathbf{B}^c \approx \mathbf{B}_2^c = \mathbf{G}_2^t \cdot \mathbf{H}_2$, where $\mathbf{G}_2$ contains the first 2 rows of $\mathbf{G} = \mathbf{D}^{1/2} \cdot \mathbf{V}_x$, and $\mathbf{H}_2$ the first 2 rows of $\mathbf{H} = \mathbf{D}^{1/2} \cdot \mathbf{V}_y$, An application biplot (Graffelman and van Eeuwijk, 2005) can be interpreted with many of the same rules as a covariance biplot for compositional data (Aitchison, 1997). The fundamental rule here is the (approximate) equivalence of the cosinus of the angle between two links with the correlation coefficient between them. Thus:

- if the link between variables A_x and B_x is parallel to the link between variables A_y and B_y, then $\operatorname{cor}[\ln(A_x/B_x), \ln(A_y/B_y)] \to \pm 1$;
- if the link between variables A_x and B_x is orthogonal to the link between variables A_y and B_y, then $\operatorname{cor}[\ln(A_x/B_x), \ln(A_y/B_y)] \to 0$.
- if the link between variables A_y and B_y is orthogonal to the link between variables C_y and D_y, then their pairwise logratios are controlled by uncorrelated factors, i.e. it is possible to change one of the ratios keeping the other fixed.

Of course, as happens in a covariance biplot, the goodness of these indications can be measured by the quality of the approximation $\mathbf{B}^c \approx \mathbf{B}_2^c$, i.e. by the fraction

$$q = \frac{d_1^2 + d_2^2}{\sum_{i=1}^{R} d_i^2}. \tag{7}$$

Thus if all singular values $d_i \approx 0, i > 2$, then the biplot perfectly describes the behaviour of $\hat{\mathbf{Y}}$. Then, for instance, if two collocated variables of $\mathbf{Y}$ have a zero length link, the explanatory composition $\mathbf{X}$ cannot change the ratio between those two variables. This does not mean that the ratio is constant, but its prediction as a function of $\mathbf{X}$ is constant.

In a practical situation, these rules may guide the analyst in the choice of relevant hypotheses of subcompositional independence between $\mathbf{X}$ and $\mathbf{Y}$, or on spliting $\mathbf{Y}$ into two subcompositions, one controlled by $\mathbf{X}$ and one independent of $\mathbf{X}$ (i.e. that formed by variables with very short rays).

4.3 Global tests of reduction of dimensionality

Later, we will see how a test can be built to check that one or more pairs of directions are uncorrelated. This test will require these directions to be chosen by the analyst. This is not the case with the pairs of singular vectors, because they are derived from the data.

To test independence along a set of optimal directions, we propose to apply a Scheffé-type joint testing,

$$D = \max_{i=1}^{r} \left(\frac{\hat{\beta}_i}{\sigma_{\beta_i}} \right)^2, \tag{8}$$

where r is the number of pairs of directions being tested for a null coefficient, $\hat{\beta}_i = d_i$ is the projected coefficient associated to each pair of directions and σ_{β_i} its estimation variance, as obtained from Eq. (10). Scheffé confidence regions check that the largest (standardized) deviation is not sufficiently different zero, which implies that there is no evidence that any other correlation between the pairs of tested directions is not different from zero. To the authors knowledge, its distribution is not known, though it can be derived under the assumption that all $\sigma_{\beta_i} = 0$. A sketch of a proof follows.

Instead of deriving the singular vectors of B as used for the biplot, the idea is to maximize Eq. (8) over all possible directions. This problem can be recasted as a generalized eigenvalue problem, seeking the largest eigenvalue of $\hat{\mathbf{\Sigma}}_{\mathbf{B}}^{-1} \cdot \operatorname{vec}^t(\hat{\mathbf{B}}) \operatorname{vec}(\hat{\mathbf{B}})$, where $\operatorname{vec}(\cdot)$ vectorizes its argument stacking the columns of the matrix in a vector, which is considered a row vector for consistency with the rest of the paper. With a bit of linear algebra, it can be shown that this eigenvalue is constant whichever bases are used to represent $\hat{\mathbf{\Sigma}}_{\mathbf{B}}$ and $\hat{\mathbf{B}}$, even non-orthogonal ones. This implies that the distribution of D can only depend on N, P and Q. For large N, it approaches a χ^2 distribution with degrees of freedom $\nu = P(Q-1)$, the number of regression coefficients. However, *large* N may mean a really large number in relation to ν. Thus, for most practical situations, that distribution must be derived by Monte Carlo methods: M samples of N simulations of a $P-1$-variate and and a $Q-1$-variate, independent normal distributions are generated; for each sample a regression model is fitted, and the largest eigenvalue of $\hat{\mathbf{\Sigma}}_{\mathbf{B}}^{-1} \cdot \operatorname{vec}^t(\hat{\mathbf{B}}) \operatorname{vec}(\hat{\mathbf{B}})$ is obtained. The estimated value $\hat{D}$ obtained with the compositional data set is then compared with those M simulated eigenvalues, to decide how anomalous is $\hat{D}$ under the hypothesis of independence.

5 Tests on the coefficients

5.1 Generalities

After a global test has proven that there is some dependence between explanatory and explained compositions, further tests can be used to check which of the pairs $(\mathbf{v}_{x,i}, \mathbf{v}_{y,i})$ show significat influences. In a more general scenario, to test the dependence between the two compositions along any pair of directions $(\mathbf{w}_x, \mathbf{w}_y)$, one must project the matrix of coefficients $\hat{\mathbf{B}}^c$ and its estimation covariance $\hat{\mathbf{\Sigma}}^c_{\mathbf{B}}$ onto those directions:

$$\hat{\beta} = \mathbf{w}_x \cdot \hat{\mathbf{B}}^c \cdot \mathbf{w}_y^t, \tag{9}$$

$$\hat{\sigma}^2_\beta = \mathbf{w}_y \otimes \mathbf{w}_x \cdot \hat{\mathbf{\Sigma}}^c_{\mathbf{B}} \cdot \mathbf{w}_y^t \otimes \mathbf{w}_x^t. \tag{10}$$

If the two directions are chosen by the analyst, a standard t-test allows to check that the projected coefficient $\hat{\beta}$ is significantly different from a reference value β_0,

$$t = \frac{\hat{\beta} - \beta_0}{\hat{\sigma}_\beta} \sim t(N - \nu), \tag{11}$$

with ν the smallest number of estimated regression parameters necessary to calculate $\hat{\beta}$ and σ^2_β. In most of the cases, that test will have as null hypothesis $\beta_0 = 0$.

If the sample size N is large, one can build an approximate test on any set of vectors at the same time. Considering $p < P$ left and $q < Q$ right vectors in the rows of the matrices $\mathbf{W}_x$ and $\mathbf{W}_y$, the projected coefficient matrix and estimation covariance are then:

$$\hat{\boldsymbol{\beta}} = \mathbf{W}_x \cdot \hat{\mathbf{B}}^c \cdot \mathbf{W}_y^t, \tag{12}$$

$$\mathbf{\Sigma}_{\boldsymbol{\beta}} = \mathbf{W}_y \otimes \mathbf{W}_x \cdot \hat{\mathbf{\Sigma}}^c_{\mathbf{B}} \cdot \mathbf{W}_y^t \otimes \mathbf{W}_x^t. \tag{13}$$

Note that these expressions hold also if $\mathbf{W}_x$ and $\mathbf{W}_y$ are not bases of the clr plane. The large sample assumption allows the asymptotic approximation

$$\mathrm{vec}(\boldsymbol{\beta}) \sim \mathcal{N}^{pq}(\mathrm{vec}(\hat{\boldsymbol{\beta}}), \mathbf{\Sigma}_{\boldsymbol{\beta}}), \quad (\mathrm{vec}(\boldsymbol{\beta}) \text{ stacked by columns})$$

thus the Mahalanobis distance between $\hat{\boldsymbol{\beta}}$ and the null hypothesis value $\boldsymbol{\beta}_0$ follows

$$\mathrm{vec}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \cdot \mathbf{\Sigma}^{-1}_{\boldsymbol{\beta}} \cdot \mathrm{vec}^t(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \sim \chi^2(pq). \tag{14}$$

In the case that the sample size is not large enough, an approach through Hotelling's T^2 distribution can be taken. In this case, if one considers a basis of $\mathbb{S}^Q$ and one single direction in $\mathbb{S}^P$, (i.e. $p = 1$ and $q = Q - 1$), then

$$F = \frac{N - P - Q}{(N - P)(Q - 1)} \hat{\boldsymbol{\beta}} \cdot \mathbf{\Sigma}^{-1}_{\boldsymbol{\beta}} \cdot \hat{\boldsymbol{\beta}}^t \sim \mathcal{F}(Q - 1, N - P - Q) \tag{15}$$

follows a Fisher $\mathcal{F}$ distribution Fahrmeir et al. (1996) under the hypothesis that $\boldsymbol{\beta}_0 = \mathbf{0}$. This test implies the complete elimination of the direction $\mathbf{W}_x$ from the problem, as it is tested whether they influence any direction of $\mathbb{S}^Q$ at all. If $q \leq Q - 1$, no exact test is known to the authors.

5.2 Testing pair-to-pair independence

The preceding tests aimed at checking the possibility to reduce the dimensionality of the linear application B, which occurs along its left and right singular vectors. However, singular vectors might no be easily interpretable in terms of the original components. For these, it would be more practical to have tests that checked that a certain subcomposition in the image space is unrelated to a certain subcomposition in the origin space.

Let us first consider the case that both subcompositions have two components, that is, to consider that one pair logratio (in the image space) is independent of another pair logratio (in the origin space). Eqs. (9) and (10) are valid whichever pair of vectors $(\mathbf{w}_x, \mathbf{w}_y)$, thus they can be of the form $\mathbf{w}_{ij}$ given in section 2. A test for independence between the subcompositions (i, j) in the origin space and (k, l) in the image space will thus be obtained by checking that the two directions $\mathbf{w}_{x,ij}$ and $\mathbf{w}_{y,kl}$ have a projected coefficient $\beta_0 = 0$. In this case, an exact distribution can be derived from Eq. (11), with the number of estimated parameters $\nu = P - 1 + 1 = P$, the number of slopes plus intercept needed to predict one single pairwise logratio in the response.

5.3 Kinds of subcompositional independence

This kind of test only assesses how the ratio between components k and l behaves as a function of the independent variable (which, in this particular case, happens to be a log-ratio of another composition). Even though such internal independence occurs, it could still happen that the explanatory variable influences the individual amounts of components k and l, in such a way that their ratio remains constant. The same concept of *internal independence* can be defined for a subcomposition in the image space: a subcomposition is said to be internally independent of an explanatory variable if all logratios between components of this subcomposition are independent of the explanatory variable.

There is a second kind of stronger subcompositional independence, namely *external independence*. A subcomposition is said to be externally independent of an explanatory variable if, apart of being internally independent, the balance between that subcomposition and the rest of the components of the image simplex is also independent of the explanatory variable. One can also then say that the dependence is restricted to that complementary subcomposition. It should be nevertheless stressed that external independence does not mean that components k or l are unrelated to the explanatory variable checked: it is just proven that both their ratio and their balance to the rest of the components of this particular composition is uncorrelated with the explanatory variable.

With regard to the origin space, the generally relevant question is the possibility to completely remove a subcomposition, i.e. restrict the explanatory power to an *explanatory subcomposition*. This can be checked forcing the internal logratios within the non-explanatory subcomposition as well as the balance between explanatory and non-explanatory components to have a zero coefficient with all response logratios.

5.4 Testing subcompositional independence

Internal subcompositional independence can be easily checked by projecting the estimated coefficients and their estimation covariance matrix onto an arbitrary basis of the subcomposition. That is, we use Eqs. (12) and (13) with $\mathbf{W}_y$ rows forming a basis of the chosen subcomposition, and $\mathbf{W}_x = \mathbf{w}_{x,ij}$ formed by one single row contrasting the two components i and j which log-ratio should be removed. External independence can be checked with the same approach, with Eqs. (12) and (13) taking in $\mathbf{W}_y$ the same rows as in the preceding case, togeher with an extra row giving the balance between the tested subcomposition and its complementary subcomposition.

Unfortunately, there is no general, analytical, exact test to check several directions of the origin space, i.e. for the cases where $\mathbf{W}_x$ has more than one row. For these cases, one needs to proceed either with approximate tests like those of Eq. (14) if the sample size is large, apply a joint Bonferroni-type testing correction to a test of the type of Eq. (11), or derive a numerical test with a Monte Carlo approach like in the case of the global test of independence. This simply requires fixing a basis of the chosen subcompositions, say with respectivel $p \leq P$ and $q \leq Q$ components, and checking that their largest generalized eigenvalue is undistinguishable from zero, with the same procedure outlined for the global independence.

5.5 Displaying the tests in diagrams

In the case of checking a 2-component subcompositional independence, the tests can be visualized with the help of confidence regions on the estimated parameters. These regions can be plotted as compositional ellipses in various standard plots for compositional data (e.g. ternary diagrams, logratio-scatterplot, etc.). In this case, we choose an explanatory direction $\mathbf{w}_x$, and take as directions in the image space $\mathbf{w}_{y,kl}$ and its orthogonal $\mathbf{w}^{\perp}_{y,kl}$, a vector with component $+1$ in the positions k and l, and $-2/(Q-2)$ in the remaining $(Q-2)$ positions. The vector $\mathbf{w}^{\perp}_{y,kl}$ represents the balance between the (k,l) subcomposition and its complementary subcomposition, denoted as $\mathbf{y}_{\overline{kl}}$. This can be displayed in a ternary diagram with vertices y_k and y_l at the base, and at the third vertex $\mathrm{g}(\mathbf{y}_{\overline{kl}})$ the geometric mean of the complementary subcomposition (fig. 1).

For a given explanatory direction $\mathbf{w}_x$, the coefficients and their estimation covariance are projected onto $\mathbf{w}_{y,kl}$ and $\mathbf{w}^{\perp}_{y,kl}$, and the adequate test, following the exact or approximate distributions mentioned in Eqs. (14) and (15), provides a typical radius of the ellipse. For instance, if the sample size is large, then the radius of an approximate test is derived from the quantiles of a chi-square distribution, and would correspond to $\sqrt{\chi^2_{0.95}(2)}$ for a test at 5% confidence. If the sample size is small, then a radius from a Hotelling's T^2 distribution is preferred, following Eq. (15), though then one tests the complete elimination of each explanatory variable, i.e. restricted explanatory power. If that ellipse contains the barycenter of the ternary diagram, then external independence can be

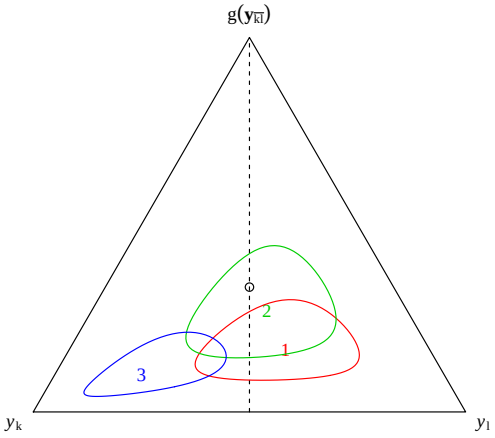


Figure 1: Example of confidence ellipses on a particular 2-part subcomposition. Ellipse 1 (red) represents a parameter estimate indicating internal independence. Ellipse 2 (green) represents a parameter estimate indicating external independence. Ellipse 3 (blue) represents a parameter estimate significantly different from both kinds of independence. The circle in the center of the diagram is the barycenter.

accepted. If it only intersects the vertical height, that corresponds to a relation $y_k : y_l = 1 : 1$, thus implying that no change in the ratio y_k/y_l can occur due to the explanatory variable, and internal independence can be accepted. Fig. 1 portrays these situations.

The same sort of display can be used to visualize a restriction in the explanatory composition. For that case, all possible pairwise logratios of the response must be displayed at the same time, and the target subcomposition to be removed must contain the barycenter of each possible pairwise ternary diagram. The regions in that case must be drawn with radii extracted from the generalized eigenvalue test, and they must be interpreted as a whole. I.e., if a certain subcomposition is checked, it suffices that one single ellipse of the set does not contain the barycenter to reject the hypothesis for the whole tested subcomposition.

The several test displays proposed in this contributions are unfortunately not fully compatible, in the sense that the ellipses of the confidence regions must be built with different radii depending on which hypothesis of subcompositional independence is tested. Ideally, an ad-hoc visualization software would be required to allow for a structured and user-friendlier exploration and visual testing of subcompositional independence. The following rules are an attempt at structuring these:

1. the sample size is large enough: build ellipses with radius $\sqrt{\chi^2_{1-\alpha}(2)}$, at the desired α confidence level, following the approximate test of Eq. (14); interpret each ellipse independently;
2. the sample size is not large enough:
   - a single pair of pre-established directions are being tested (e.g. a pairwise logratio in each origin and image space); use the exact test of Eq. (11); this can also be used to derive a radius following a Fisher F distribution:
     - for internal independence only, use 1 and $N - P$ degrees of freedom;
     - for both internal and external independence only, use 2 and $N-P$ degrees of freedom, though this is not an exact test, it can be taken as an approximate visualization;
     - for restriction of explanatory power, use $Q - 1$ and $N - P - Q$) degrees of freedom; this is an exact test if used to eliminate completely an explanatory variable from the problem, i.e. the barycenter of each ternary diagram should be contained in the resulting ellipse;
   - a pair of subcompositions are being tested; use the exact test of Eq. (8); the test provides the largest $D_{1-\alpha}$ admissible for an α confidence level, if λ_1 represents the largest eigen-

value of $\mathbf{\Sigma}_{\boldsymbol{\beta}}^{-1} \cdot \mathrm{vec}(\boldsymbol{\beta})^t \mathrm{vec}(\boldsymbol{\beta})$, with $\mathbf{\Sigma}_{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ projected onto the chosen subcompositions through Eqs.(12) and (13), then the radius of the ellipses is $D_{1-\alpha}/\lambda_1$; this visualization corresponds to an exact test if the resulting ellipses contain the barycenter on all ternary diagrams involved.

6 Case studies

6.1 Predicting magnetic separation results

Production in a mine follows many steps from extraction of rock to the final dispatch of refined material. Several of these processes aim at concentrating the value elements. Among them, magnetic separation is used to split ferromagnetic material from non-ferromagnetic ones. To evaluate feasibility of mining Lithium from Zinnwald, Saxony, Germany, experiments of magnetic separation of Li/Fe/K-rich sheet silicates have been conducted. Three rock samples have been separately milled and classified into 5 different particle sizes. Each of the resulting 15 samples was analysed for the composition Li-Fe-K-Al, and then it was divided in 8 aliquots. Each was applied a magnetic separation with 8 different intensities of magnetic susceptibility, giving rise to 120 observations. The final composition in the same 4 components was analysed. The goal is to quantify the degree of control the input composition has on the output composition, specially with regard to Li.

	Li.o	Fe.o	K.o	Al.o
Li.i	0.04	-0.28	-0.04	0.27
Fe.i	2.72	1.37	-0.65	-3.43
K.i	-2.79	-1.04	0.65	3.18
Al.i	0.03	-0.06	0.05	-0.02

Table 1: Estimated matrix of clr regression coefficients relating input and output composition before and after a magnetic separation process.

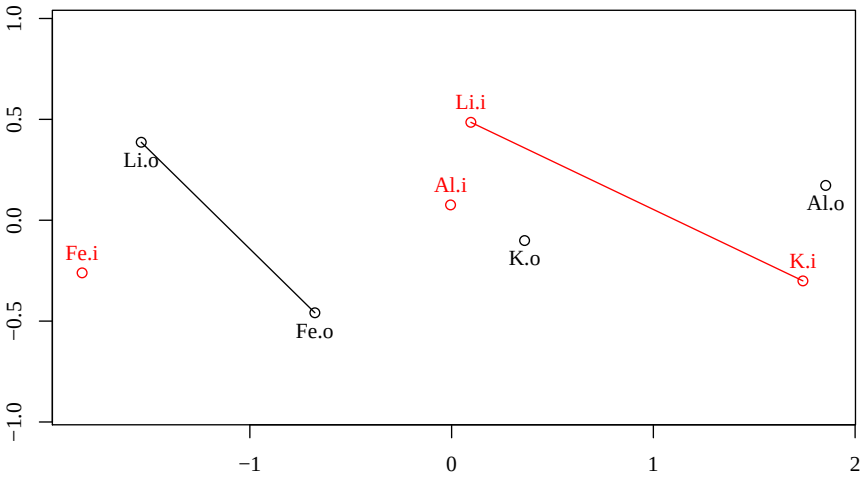


Figure 2: Biplot of the linear application relating input and output composition before and after a magnetic separation process.

For this goal, a black-box ilr was applied to both compositions, and the resulting matrix of coefficients was recasted to clr (table 1). The biplot representation of that system is shown in fig. 2.

A global test can be conducted, by taking the test of Eq. (8). This gives a p-value $< 10^{-5}$, thus we cannot accept that all eigenvalues are zero, and the linear application must have a rank larger than 0.

The biplot (fig. 2) shows that the ratio Li/Fe in the output should favor a larger Li/K in the input, as these two links are sub-parallel. This can be verified by a simple linear regression

of these two logratios, which gives a significance of the zero-slope hypothesis of 0.000602, and an $R^2 = 0.09531$. Thus, a certain connection exists, but it is rather weak. This is also visible in fig. 3, where the ellipse corresponding to input Li/K for K/Fe on the output does not contain the barycenter of the diagram. In the biplot one can also see that the input ratio Al/Fe controls all internal ratios between Fe, K and Al at the output. This corresponds to the cyan ellipses in the ternary diagrams of fig. 3, which do barely intersect the vertical height for Fe/K and Al/K, but does not for Al/Fe (i.e. this last shows a significant relation).

These results suggest that the input subcomposition {Li,K} does not influence the output subcomposition {K,Al,Fe}, while input {Fe,Al} does not condition output {Li,Fe}. These have been tested with Eq. (8) again, giving respectively p-values of 0.50 and 0.22 respectively. Though these tests have been conducted separately, these large p-values support accepting both subcompositional independence hypotheses, i.e. {K,Al,Fe} is internally independent of the input ratio Li/K, while {Li,Fe} is internally independent of the ratio Fe/Al. These relations are already hinted in the biplot, as the links between these pairs of subcompositions are roughly mutually orthogonal.

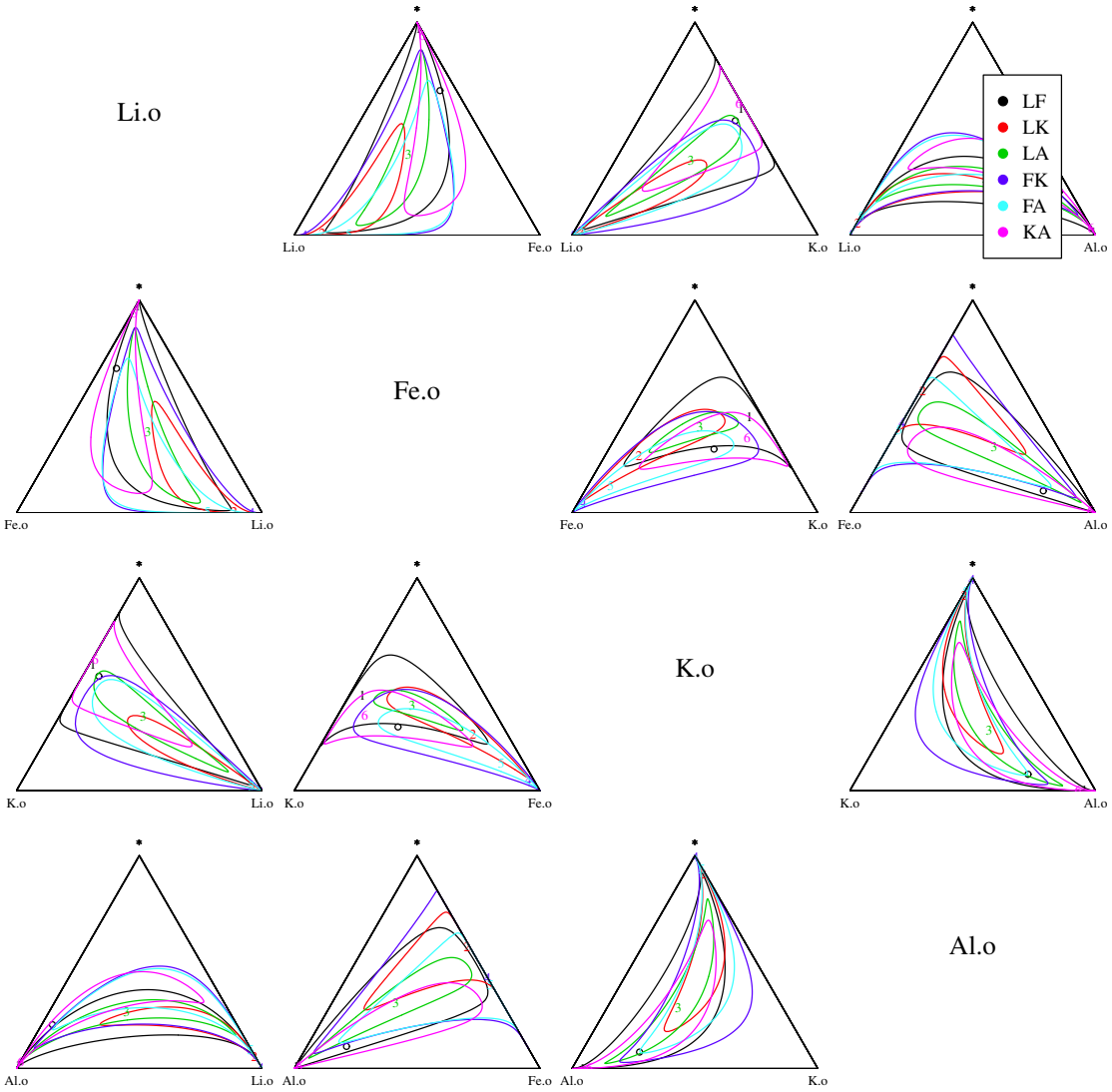


Figure 3: Confidence ellipses at a 95% confidence on each possible pairwise logratio, built with a radius derived from Hotelling's T^2 test.

A further study of fig. 3 shows that the ellipses for all explanatory ratios except Li/K do always contain the barycenter. That implies that beyond this rattio, the input variables have no

explanatory power whatsoever. This can be checked by computing the following two models:

$$\begin{aligned}\mathrm{ilr}(\hat{\mathbf{Y}}) &= \mathbf{b}_0 + \mathbf{b}_1 \cdot \ln \frac{K}{Al \cdot Fe \cdot Li} + \mathbf{b}_2 \cdot \ln \frac{Li}{Al \cdot Fe} + \mathbf{b}_3 \cdot \ln \frac{Al}{Fe},\\ \mathrm{ilr}(\hat{\mathbf{Y}}') &= \mathbf{b}'_0 + \mathbf{b}'_3 \cdot \ln \frac{Li}{K},\end{aligned}$$

and comparing the residual variances between the two systems, for instance with a $\chi^2(6)$-test of an ANOVA table. That gives a p-value of 0.26, showing that the removal of the first and second ilr input coordinate is justified. In a similar way, the figure shows that the output ratio K/Al cannot be predicted, because all ellipses in this ternary diagram cross the vertical height.

6.2 Relations between educational level and political preferences

The following example relates the proportions of votes of each of the five major parties in the Autonomous Region of Catalunya (Spain) with the educational level of the population. The region is divided in shires, which serve both as statistical districts for social surveys and as election districts. In each of the registered elections (to the autonomous Parliament sitting in Barcelona), the number of votes to each of the following 5 parties was recorded: PP (conservative-liberal, present in all Spain), PSC (social-democratic, allied with the Socialist Party of Spain), CiU (liberal-conservative, catalan pro-autonomy), ERC (social-democratic, catalan pro-independence) and IC (communist-green coalition, present in all Spain, but essentially in the big cities). Results cover the election years 1980, 1984, 1988, 1992, 1995, 1999, 2003 and 2006. With regard to educational level, each shire population was classified in 6 levels, according to the highest level obtained: uneducated/illiterate (0s), primary school (1s), secondary school (2s), professionalizing/vocational education (P), university degree (U) and university master/PhD (M). Statistics with regard to educational level were compiled for the years 1981, 1986, 1991, 1996 and 2001, and show a clear trend towards higher educational level. Each election year was associated with the nearest year where educational stats were available.

	CiU	PSC	PP	IC	ERC
s0	0.1712	0.2465	-0.5587	0.3865	-0.2454
s1	0.1914	-0.5428	1.4117	-1.0534	-0.0069
s2	0.2098	0.4996	-0.4199	0.0825	-0.3720
P	-0.3531	0.0653	-1.6347	1.0987	0.8237
U	0.2409	0.3182	-0.0345	-0.3950	-0.1296
M	-0.4601	-0.5868	1.2361	-0.1193	-0.0698

Table 2: Estimated matrix of clr regression coefficients relating election results with educational level.

For each shire and each election year, the 5-part composition of votes was related to the 6-part composition of educational level. Using a black-box ilr and expressing the results in clr, we obtain the matrix of coefficients given in Table 2. The first two pairs of singular vectors of this matrix capture a $q = 0.957$ fraction of the explained variance of the election datasets (Eq. 7). It may be thus assumed that most of the dependence between the two compositions is captured by the biplot they define (fig. 4). An analysis of this biplot shows several pairs of variables that might show good correlations (fig. 5), though they are indeed not very good.

Finally, confidence ellipses (fig. 6) obtained with Hotelling's T^2 distribution show:

- In the diagram PSC-CiU, no ellipse involving s1 crosses the vertical height, thus all ratios involving s1 are significant. Actually ratios with s1 in the numerator fall near to CiU, while ratios involving it in the denominator fall nearer to PSC. Thus, CiU/PSC significantly increases with increasing s1 proportion.
- A similar picture can be seen in the diagram ERC-IC, where significant ellipses involve s1 against s0, s2, M and P, and in all cases increasing s1 favors ERC against IC. Note that a similar conclusion can be extracted from the biplot (fig. 4), as the links pointing towards s1 from s0, s2 and P would have a positive projection on the link IC-ERC.
- The PP-CiU diagram show that the ratios s1/s2 and s1/P favor PP, while the ratios P/M, s0/s1 and s0/M favor CiU. These again show a certain consistent pattern that decreasing s1 favors CiU against PP.

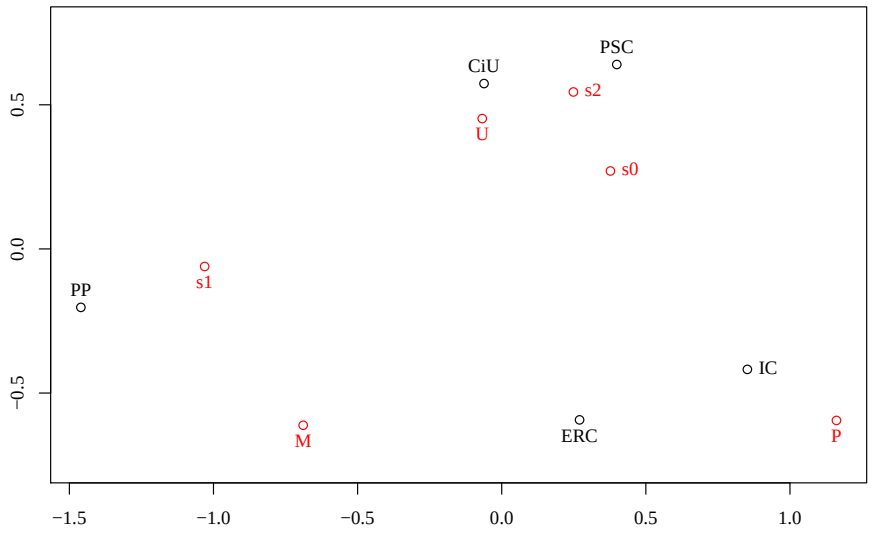


Figure 4: Biplot of the linear application explaining election results as a function of educational level.

- In the PP-PSC diagram we see again significant effects for all ratios involving s1 in such a way that the ratio PP/PSC increases, but also most of the dark blue ellipses (involing M in the denominator) favor PP/PSC, thus the conservative party is benefited in shires with dominance of s1 and M against the other educational levels. Interstingly, the ratio M/s1 does not affect the PSC/PP distribution, as both would tend to favor PP / disfavor PSC. This is consistent with the biplot, as M and s1 fall near to PP while U, s0 and s2 fall near the PSC.

7 Conclusions

Linear regression models can be built between two compositional data sets, by means of the log-ratio approach: each of the two involved compositions is expressed in its own one-to-one logratios, and the resulting scores are linked through a multivariate multiple linear regression model. Classical least-squares fitting techniques provide estimates of the intercept and all slopes, as well as their estimation covariance. These elements can be interpreted in compositional terms, the intercept as a composition in the image space, the slopes all together as a linear application between the origin and the image spaces. Note that all these results are not dependent on the sets of logratios actually used to obtain the coefficients or the variance matrix; they identify intrinsic objects, with an own geometric interpretation.

The difficulties of a composition-to-composition regression stem from the large dimension of the parameter space, and the lack of an easy interpretation of each estimated coefficient in terms of the original components. In this contribution, three kinds of independence hypotheses have been presented, all testable in terms of logratios and still interpretable in terms of the original parts: internal subcompositional independence (involving only ratios within an explained subcomposition), external independence (extending internal independence to include also a balance between the chosen subcomposition and the remaining parts), and finally restricted explanatory power (recyprocal to external independence, but involving an explanatory subcomposition).

Some of these hypotheses can be tested with existing tests (exact Hotelling's T^2 or approximate χ^2-tests). However, we have also presented an omnibus test, similar to Scheffé joint testing in ANOVA applications. This is based on a generalized eigenvalue problem of the matrix of regression coefficients and their estimation variance-covariance matrix. This can be applied to test global independence, as well as any of the mentioned subcompositional independence hypotheses.

Finally, several graphical representations are available to visually display the model and these tests. Parallel to redundancy analysis, a biplot representation of the linear application linking both compositions can be used as an exploratory tool. All tests presented here can also be represented as confidence ellipses on ternary diagrams. Each hypothesis of independence is represented as a line (a height of the ternary diagram) or a point (the barycenter). If a confidence ellipse contains/intersects these geometric objects, then the corresponding hypothesis cannot be rejected.

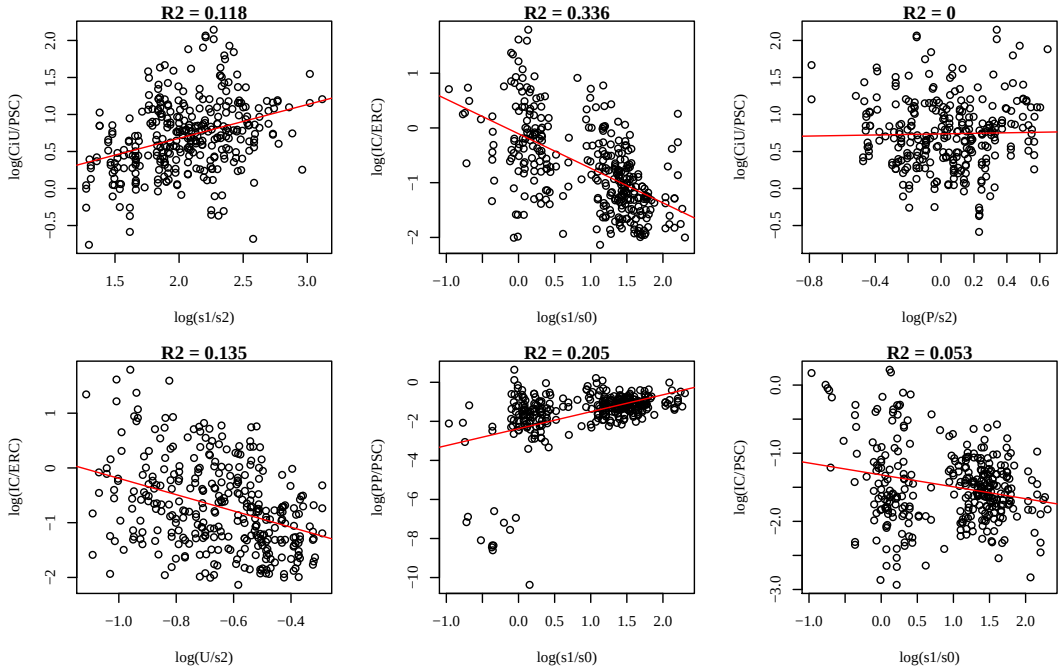


Figure 5: Scatterplots of selected log-ratio proportions between votes and educational level. The first 4 were selected because of their parallel links (implying a good relation). The two from the right column were selected because of their orthogonal links (implying uncorrelation). An R^2 goodness of fit of the linear regression is shown for each figure.

These tools and concepts ease the structured, systematic characterization of a regression model between two compositions. The complexity of the problem requires nevertheless further developments, like e.g. some interactive software facilities. This is left for future work.

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. (1997). The one-hour course in compositional data analysis or compositional data analysis is simple. In V. Pawlowsky-Glahn (Ed.), *Proceedings of IAMG'97 - The III Annual Conference of the International Association for Mathematical Geology*, Volume I, II and addendum, Barcelona (E), pp. 3–35. International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E), 1100 p.

Aitchison, J. (2002). Simplicial inference. In M. A. Viana and D. S. Richards (Eds.), *Algebraic Methods in Statistics and Probability*, Volume 287 of *Contemporary Mathematics (American Mathematical Society)*, pp. 1–22. University of Notre Dame, Notre Dame, Indiana: American Mathematical Society, Providence, Rhode Island (USA), 340 p.

Eaton, M. L. (1983). *Multivariate Statistics. A Vector Space Approach*. John Wiley and Sons.

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

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.

Fahrmeir, L., A. Hamerle, and G. Tutz (1996). *Multivariate statistische Verfahren*.

Gabriel, K. R. (1971). The biplot – graphic display of matrices with application to principal component analysis. *Biometrika 58*(3), 453–467.

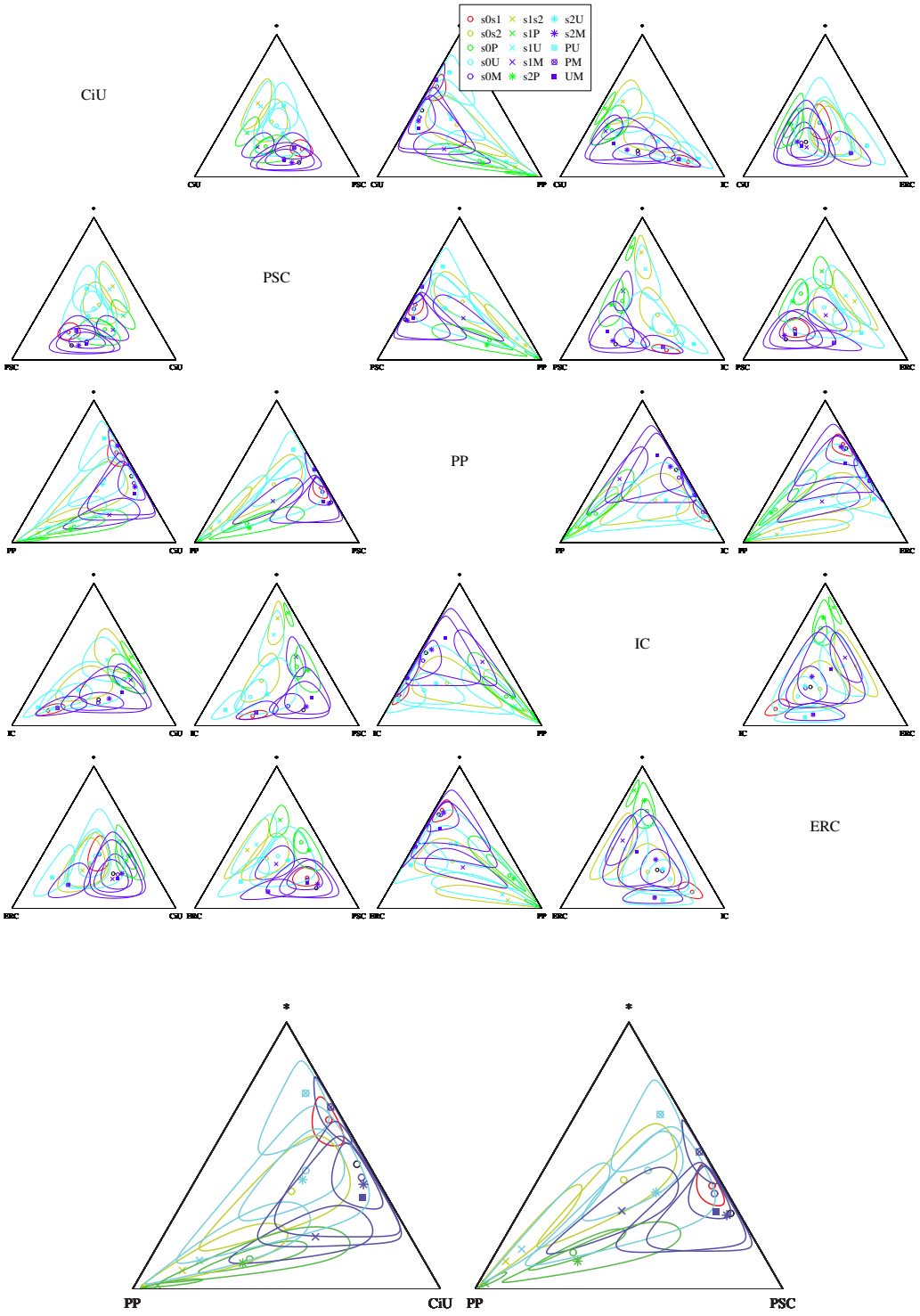


Figure 6: Hotelling's T^2 5%-confidence ellipses for the dependence of election results on educational level. Note that the figure is symmetric. Color depends on the numerator of the ratio on the explanatory variable while symbol depends on the denominator of that ratio. Legend descriptors give first the denominator, then the numerator, e.g. s0s1 means $\log(s0/s1)$. The two lower plots show a zoom of the diagrams PP-CiU and PP-PSC

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

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.

Local regression for compositional data

C. VENIERI[1] and M. DI MARZIO[1] and A. PANZERA[1]

[1]SEQFE Department - University "G. d'Annunzio" of Chieti-Pescara, Italy c.venieri@unich.it

Abstract

Regression for compositional data has been considered only from a parametric point of view. We introduce a nonparametric estimator for the regression function when both the response variable and the predictor can be defined on the simplex, and derive its asymptotic properties. To this end, we use the isometric log-ratio transformation along with properly defined kernel functions. The performances of our estimator are compared to those of a parametric model using a real data set.

1 Introduction

Regression for compositional data have been studied with real covariates and a compositional response or viceversa, with the most recent contributions provided by Tolosana-Delgado and Van Den Boogart (2011) and Egozcue et al. (2012), which worked with orthonormal coordinates. Regression models, where the Dirichlet or the logistic-normal densities were considered as distribution of residuals, have been tackled, for example, by Aitchison (1982) and Hijazi and Jernigan (2009). All of the proposed approaches have parametric nature, whereas nonparametric methods for regression are totally unexplored for compositional data. The purpose of this work is to introduce nonparametric regression estimators when both predictor and response variables take values on the simplex, or only one of them is compositional and the other one is real. In particular, we propose kernel estimators for these kinds of regression.

Kernel regression is aimed to estimate the conditional mean of the response variable, which is not assumed to belong to a known parametric family, only a few smoothness assumptions being required. The estimate of the regression function at a point is carried out by locally weighted mean of the observations of the response. The local weights are defined by means of functions, called kernels, rescaled by a factor $h > 0$. These weights give a greater contribution, in the estimation, to the observations which are closest to the estimation point. The factor h determines the spread of the kernel and its choice is crucial for the performance of the estimator. See, for example, Härdle (1990) and Wand and Jones (1995) for classical references on the subject.

Kernel regression with a compositional predictor and a real-valued response entails the adaptations of the standard theory, throughout the definition of suitable weights. Differently, when the response lies on the simplex, the target function turns out to be defined as the conditional centre of a random composition, and a suitable defined local *average* has to be considered. In pursuing our nonparametric approach, we apply the principle of *working on coordinates*, using the isometric log-ratio transformation (see Egozcue et al. (2003) for details).

This work is organized as follows. In Section 2, after discussing some weight functions, we define a kernel estimator for the case of a real-valued response and a compositional predictor, and derive its asymptotic properties. The estimator for the case of a compositional response, along with its asymptotic properties, is the subject of sections 3 and 4 for the case of compositional and real predictor, respectively. Finally, Section 5 is devoted to illustrate a real case study.

2 Simplicial-real regression

When the predictor is compositional and the response is real, the kernel smoothing method requires the definition of suitable weight functions. Specifically, to take into account the compositional nature of the predictor, we need to consider kernels defined on the simplex. Aitchison and Lauder (1985), for the task of density estimation, used the Dirichlet kernel and the additive logistic normal one, both suffering from some drawbacks, such as the necessity of choosing a single smoothing parameter in order to guarantee invariance, while Chacón et al. (2011) used the normal kernel as defined in Mateu-Figueras et al. (2003). This latter is a density on the simplex $\mathcal{S}^D$ with respect to the Aitchison measure, while, in the coordinate space $\mathbb{R}^{D-1}$, it is a density with respect to the Lebesgue measure.

In a similar way, we can define several kernels on $\mathcal{S}^D$, or simplicial kernels, as standard euclidean kernels defined on the space of coordinates, which are densities on $\mathcal{S}^D$ with respect to the Aitchison

measure. For example, the simplicial Epanechnikov kernel can be defined as

$$K(\mathbf{x}) = \frac{D+1}{2C_{D-1}} \left[1 - \text{ilr}(\mathbf{x})^\top \text{ilr}(\mathbf{x})\right] 1_{\{\text{ilr}(\mathbf{x})^\top \text{ilr}(\mathbf{x})<1\}}$$

where C_D is the volume of the unitary D-dimensional sphere, $\text{ilr}(\mathbf{x})$ is the isometric log-ratio transformation of $\mathbf{x}$, while $1_{\{A\}}$ stands for the indicator function of the set A.

Centering the kernel on the observation $\mathbf{X}_i$, $i \in \{1, \ldots, n\}$, and rescaling by a $(D-1) \times (D-1)$ positive definite smoothing matrix $\mathbf{H}$, we obtain

$$K_{\mathbf{H}}(\mathbf{x} \ominus \mathbf{X}_i) = \frac{D+1}{2C_{D-1}} |\mathbf{H}|^{-1} \left[1 - \text{ilr}(\mathbf{x} \ominus \mathbf{X}_i)^\top (\mathbf{H}\mathbf{H}^\top)^{-1} \text{ilr}(\mathbf{x} \ominus \mathbf{X}_i)\right] 1_{\{\text{ilr}(\mathbf{x} \ominus \mathbf{X}_i)^\top (\mathbf{H}\mathbf{H}^\top)^{-1} \text{ilr}(\mathbf{x} \ominus \mathbf{X}_i)<1\}}$$

where, for any $\mathbf{a}$ and $\mathbf{b} \in \mathcal{S}^D$, $\mathbf{a} \ominus \mathbf{b} = \mathcal{C}\left(\frac{a_1}{b_1}, \ldots, \frac{a_D}{b_D}\right)$, with $\mathcal{C}(\mathbf{u}) = \left(\frac{u_1}{\sum_{i=1}^D u_i}, \ldots, \frac{u_D}{\sum_{i=1}^D u_i}\right)$ being the closure of any vector $\mathbf{u} \in \mathbb{R}_+^D$.

The first and second moments of a simplicial kernel are, respectively

$$\boldsymbol{\mu}_1(K) := \int_{\mathbb{R}^{D-1}} K(\mathbf{x}) \text{ilr}(\mathbf{x}) d\lambda(\text{ilr}(\mathbf{x})) = \mathbf{0}_{D-1},$$

and

$$\mu_2(K)\mathbf{I}_{D-1} := \int_{\mathbb{R}^{D-1}} K(\mathbf{x}) \text{ilr}(\mathbf{x}) (\text{ilr}(\mathbf{x}))^\top d\lambda(\text{ilr}(\mathbf{x})),$$

where λ denotes the Lebesgue measure on $\mathbb{R}^{D-1}$, $\mathbf{0}_{D-1}$ and $\mathbf{I}_{D-1}$ stand for the $D-1$ dimensional zero vector, and the identity matrix of order $D-1$, respectively.

Now, given the $\mathcal{S}^D \times \mathbb{R}$-valued random vector $(\mathbf{X}, Y)$, the dependence of Y from $\mathbf{X}$ is suitably described by the function $m : \mathcal{S}^D \to \mathbb{R}$ minimizing $\text{E}[(Y - m(\mathbf{X}))^2 | \mathbf{X}]$. It is well known that this function, which we call simplicial-real regression function, at $\mathbf{x} \in \mathcal{S}^D$, is defined by $m(\mathbf{x}) = \text{E}[Y | \mathbf{X} = \mathbf{x}]$.

Now, given the $\mathcal{S}^D \times \mathbb{R}$-valued random sample $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$, we assume the model

$$Y_i = m(\mathbf{X}_i) + \epsilon_i, \quad i = 1, \ldots, n,$$

where the ϵ_is, conditioned on the $\mathbf{X}_i$s, are independent and identically distributed (*i.i.d.*) $\mathbb{R}$-valued random variables, satisfying $\text{E}[\epsilon_i | \mathbf{X}_i] = 0$, $\text{Var}[\epsilon_i | \mathbf{X}_i] = \sigma^2(\mathbf{X}_i) < \infty$, and are independent from the $\mathbf{X}_i$s.

Hence, a kernel estimator for m at $\mathbf{x} \in \mathcal{S}^D$ can be defined as

$$\hat{m}(\mathbf{x}; \mathbf{H}) := \frac{\sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} \ominus \mathbf{X}_i) Y_i}{\sum_{i=1}^n K_{\mathbf{H}}(\mathbf{x} \ominus \mathbf{X}_i)}, \tag{1}$$

where K is a simplicial kernel.

In what follows we set $\mathbf{u}^* = \text{ilr}(\mathbf{u})$, for any $\mathbf{u} \in \mathcal{S}^D$, and, for any function $g : \mathcal{S}^D \to \mathbb{R}$, we define $\tilde{g} : \mathbb{R}^{D-1} \to \mathbb{R}$ such that $\tilde{g} := g \circ \text{ilr}^{-1}$. Notice that, since a simplicial kernel depends on $\mathbf{x} \in \mathcal{S}^D$ throughout $\text{ilr}(\mathbf{x})$, it holds that $K(\mathbf{x}) = \tilde{K}(\mathbf{x}^*)$, for each $\mathbf{x} \in \mathcal{S}^D$.

To derive asymptotic properties for estimator (1), we start by observing that, letting p denote the density function of a random variable $\mathbf{X}$ supported on $\mathcal{S}^D$, for a function $g : \mathcal{S}^D \to \mathbb{R}$, it holds that

$$\int_{\mathcal{S}^D} g(\mathbf{x}) p(\mathbf{x}) d\lambda_a(\mathbf{x}) = \int_{\mathbb{R}^{D-1}} \tilde{g}(\mathbf{x}^*) \tilde{p}(\mathbf{x}^*) d\mathbf{x}^*,$$

where λ_a stands for the Aitchison measure on the simplex. Hence, denoting the design density as f, after some standard calculations within the coordinates space, we are able to state

Result 1 *Given a random sample $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ taking values on $\mathcal{S}^D \times \mathbb{R}$, consider the estimator (1). If*

a) $\tilde{f}$ is differentiable, and all second derivatives of $\tilde{m}$ are continuous,

b) $\mu_2(K) \neq 0$,

c) each entry of $\mathbf{H}$ and $(n|\mathbf{H}|)^{-1}$ tend to 0 as $n \to \infty$,

then

$$\begin{aligned}&\mathrm{E}[\hat{m}(\mathbf{x};\mathbf{H}) - m(\mathbf{x}) \mid \mathbf{X}_1,\ldots,\mathbf{X}_n]\\ &= \mu_2(K)\left(\frac{1}{2}\mathrm{tr}\left\{\mathbf{H}^\top \mathcal{H}_{\tilde{m}}(\mathbf{x}^*)\mathbf{H}\right\} + \frac{D_{\tilde{m}}(\mathbf{x}^*)^\top \mathbf{H}\mathbf{H}^\top D_{\tilde{f}}(\mathbf{x}^*)}{\tilde{f}(\mathbf{x}^*)}\right) + o(\mathrm{tr}\{\mathbf{H}\mathbf{H}^\top\}),\end{aligned}$$

where, for a whatever function g, $D_g(u^*)$ and $\mathcal{H}_g(u^*)$ denote the Jacobian vector and the Hessian matrix of g at u^* respectively, and

$$\mathrm{Var}[\hat{m}(\mathbf{x};\mathbf{H}) \mid \mathbf{X}_1,\ldots,\mathbf{X}_n] = \frac{\tilde{\sigma}^2(\mathbf{x}^*)}{n|\mathbf{H}|\tilde{f}(\mathbf{x}^*)}\int\{\tilde{K}(\mathbf{z}^*)\}^2 d\mathbf{z}^* + o\left(\frac{1}{n|\mathbf{H}|}\right).$$

As a criterion to obtain the optimal amount of smoothing, we consider the asymptotic version of the conditional mean integrated squared error $\mathrm{MISE}[\hat{m}(\cdot;\mathbf{H}) \mid \mathbf{X}_1,\ldots,\mathbf{X}_n]$, which is the sum of the integrated conditional squared bias and the integrated conditional variance of $\hat{m}(\mathbf{x};\mathbf{H})$. For the simplest case where $\mathbf{H} = h\mathbf{I}_{D-1}$, based on Result 1, we have that the value of h minimizing the leading term of $\mathrm{MISE}[\hat{m}(\cdot;\mathbf{H}) \mid \mathbf{X}_1,\ldots,\mathbf{X}_n]$ is

$$h_{\mathrm{opt}} = \left\{\frac{(D-1)V}{4W}\right\}^{\frac{1}{D+3}} n^{-\frac{1}{D+3}} \tag{2}$$

where V (W respectively) is the part of the leading term of the integrated variance (squared bias resp.) of $\hat{m}$ not depending on h and n.

3 Simplicial-simplicial regression

Given the random variables $\mathbf{X} \in \mathcal{S}^D$ and $\mathbf{Y} \in \mathcal{S}^L$, the dependence of $\mathbf{Y}$ on $\mathbf{X}$ could be described by the function $\mathbf{m} : \mathcal{S}^D \to \mathcal{S}^L$ which minimizes $\mathrm{E}[d_a^2(\mathbf{Y}, \mathbf{m}(\mathbf{X}))|\mathbf{X} = \mathbf{x}]$, where

$$d_a(\mathbf{x},\mathbf{y}) = \sqrt{\frac{1}{D}\sum_{i<j}\left\{\ln\left(\frac{x_i}{x_j}\right) - \ln\left(\frac{y_i}{y_j}\right)\right\}^2}$$

is the Aitchison distance, which satisfies $d_a(\mathbf{x},\mathbf{y}) = d(\mathbf{x}^*,\mathbf{y}^*)$, where $d(\mathbf{a},\mathbf{b})$ stands for the euclidean distance between vectors $\mathbf{a}$ and $\mathbf{b}$. The function $\mathbf{m}$, which we call simplicial-simplicial regression function, defines the conditional centre of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ and can be expressed as

$$\begin{aligned}\mathbf{m}(\mathbf{x}) &= \mathrm{ilr}^{-1}(\mathrm{E}[\mathrm{ilr}(\mathbf{Y})|\mathbf{X} = \mathbf{x}])\\ &= \mathrm{ilr}^{-1}(\mathrm{E}[\mathrm{ilr}_1(\mathbf{Y})|\mathbf{X} = \mathbf{x}],\ldots,\mathrm{E}[\mathrm{ilr}_{L-1}(\mathbf{Y})|\mathbf{X} = \mathbf{x}]).\end{aligned} \tag{3}$$

Notice that the argument of the above ilr inverse is $\mathbf{m}^*(\mathbf{x}^*) = (m_1^*(\mathbf{x}^*),\ldots,m_{L-1}^*(\mathbf{x}^*))^\top$. In particular, letting $\mathbf{U}$ and $\mathbf{V}$ be the basis contrast-matrices of two orthonormal bases for $\mathcal{S}^D$ and $\mathcal{S}^L$ respectively (see Egozcue et al. (2011) for a definition of basis contrast-matrix), and $\mathrm{ilr}_{\mathbf{U}}$ and $\mathrm{ilr}_{\mathbf{V}}$ the corresponding ilr transformations, then $\mathbf{m}^* = \mathrm{ilr}_{\mathbf{V}} \circ \mathbf{m} \circ \mathrm{ilr}_{\mathbf{U}}^{-1}$ is a function from $\mathbb{R}^{D-1}$ to $\mathbb{R}^{L-1}$, as shown in the figure below.

$$\begin{array}{ccc}\mathcal{S}^D & \xrightarrow{\mathbf{m}} & \mathcal{S}^L\\ {\scriptstyle \mathrm{ilr}_{\mathbf{U}}}\downarrow & & \downarrow{\scriptstyle \mathrm{ilr}_{\mathbf{V}}}\\ \mathbb{R}^{D-1} & \xrightarrow[\mathbf{m}^*]{} & \mathbb{R}^{L-1}\end{array}$$

Since $d_a(\mathbf{Y},\mathbf{m}(\mathbf{x})) = d^2(\mathbf{Y}^*,\mathbf{m}^*(\mathbf{x}^*))$, if $\mathbf{m}$ is the function which minimizes $\mathrm{E}[d_a^2(\mathbf{Y},\mathbf{m}(\mathbf{X}))|\mathbf{X}]$, then $\mathbf{m}^*$ is the function such that $\mathrm{E}[d^2(\mathbf{Y}^*,\mathbf{m}^*(\mathbf{X}^*))|\mathbf{X}^*]$ is minimized, too (see Pawlowsky-Glahn and Egozcue (2001)), *i.e.* $\mathbf{m}^*$ is the regression function of $\mathrm{ilr}_{\mathbf{V}}(\mathbf{Y})$ on $\mathrm{ilr}_{\mathbf{U}}(\mathbf{X})$.

Given a random sample $(\mathbf{X}_1,\mathbf{Y}_1),\ldots,(\mathbf{X}_n,\mathbf{Y}_n)$, taking values on $\mathcal{S}^D \times \mathcal{S}^L$, we assume the model

$$\mathbf{Y}_i = \mathbf{m}(\mathbf{X}_i) \oplus \boldsymbol{\epsilon}_i, \qquad i = 1,\ldots,n,$$

where $\oplus$ denotes *perturbation* on the simplex (see, for example, Aitchison (1986)), and the $\boldsymbol{\epsilon}_i$s, conditioned on the $\boldsymbol{X}_i$s, are *i.i.d.* compositional random variables with $\mathrm{ilr}^{-1}(\mathrm{E}[\mathrm{ilr}(\boldsymbol{\epsilon}_i)|\mathbf{X}_i]) = \mathrm{ilr}^{-1}(\mathbf{0}) = (D^{-1},\ldots,D^{-1})$ and $\mathrm{Var}[\mathrm{ilr}_j(\boldsymbol{\epsilon}_i)|\mathbf{X}_i] = \sigma_j^2(\mathbf{X}_i) < \infty$, for $j \in \{1,\ldots,L-1\}$.

We propose as an estimator of $\mathbf{m}$ at $\mathbf{x} \in \mathcal{S}^D$ the vector-valued function

$$\hat{\mathbf{m}}(\mathbf{x};\mathbf{H}) := \mathrm{ilr}^{-1}\left(\hat{m}_1^*(\mathbf{x}^*;\mathbf{H}),\ldots,\hat{m}_{L-1}^*(\mathbf{x}^*;\mathbf{H})\right) \tag{4}$$

where, for $j \in \{1,\ldots,L-1\}$,

$$\hat{m}_j^*(\mathbf{x}^*;\mathbf{H}) := \frac{\sum_{i=1}^n \tilde{K}_{\mathbf{H}}(\mathbf{X}_i^* - \mathbf{x}^*)\mathrm{ilr}_j(\mathbf{Y}_i)}{\sum_{i=1}^n \tilde{K}_{\mathbf{H}}(\mathbf{X}_i^* - \mathbf{x}^*)}. \tag{5}$$

An accuracy measure for the above estimator can be defined as

$$\begin{aligned}\mathcal{L}[\hat{\mathbf{m}}(\mathbf{x};\mathbf{H})] &:= \mathrm{E}[d_a^2(\hat{\mathbf{m}}(\mathbf{x};\mathbf{H}),\mathbf{m}(\mathbf{x}))\mid \mathbf{X}_1,\ldots,\mathbf{X}_n] \\ &= \mathrm{E}[d^2(\hat{\mathbf{m}}^*(\mathbf{x}^*;\mathbf{H}),\mathbf{m}^*(\mathbf{x}^*))\mid \mathbf{X}_1^*,\ldots,\mathbf{X}_n^*].\end{aligned}$$

which can be decomposed as

$$\mathcal{L}[\hat{\mathbf{m}}(\mathbf{x};\mathbf{H})] = \sum_{j=1}^{L-1}\{\mathrm{E}[\hat{m}_j^*(\mathbf{x}^*;\mathbf{H})\mid \mathbf{X}_1^*\ldots,\mathbf{X}_n^*] - m_j^*(\mathbf{x}^*)\}^2 + \sum_{j=1}^{L-1}\mathrm{Var}[\hat{m}_j^*(\mathbf{x}^*;\mathbf{H})\mid \mathbf{X}_1^*\ldots,\mathbf{X}_n^*],$$

where the summands in both right hand side terms of the above equation are provided in the following

Result 2 *Given the $\mathcal{S}^D \times \mathcal{S}^L$-valued random sample $(\mathbf{X}_1,\mathbf{Y}_1),\ldots,(\mathbf{X}_n,\mathbf{Y}_n)$, consider estimator (4). If assumption a), with m_j^* in place of $\tilde{m}$, and assumptions b) − c) of Result 1 hold, then, for $j \in \{1,\ldots,L-1\}$*

$$\begin{aligned}&\mathrm{E}[\hat{m}_j^*(\mathbf{x}^*;\mathbf{H})\mid \mathbf{X}_1^*\ldots,\mathbf{X}_n^*] - m_j^*(\mathbf{x}^*) \\ &\qquad = \mu_2(K)\left(\frac{1}{2}\mathrm{tr}\left\{\mathbf{H}^\top \mathcal{H}_{m_j^*}(\mathbf{x}^*)\mathbf{H}\right\} + \frac{D_{m_j^*}(\mathbf{x}^*)^\top \mathbf{H}\mathbf{H}^\top D_{\tilde{f}}(\mathbf{x}^*)}{\tilde{f}(\mathbf{x}^*)}\right) + o(\mathrm{tr}\{\mathbf{H}\mathbf{H}^\top\})\end{aligned}$$

and

$$\mathrm{Var}[\hat{m}_j^*(\mathbf{x}^*;\mathbf{H})\mid \mathbf{X}_1^*\ldots,\mathbf{X}_n^*] = \frac{\tilde{\sigma}_j^2(\mathbf{x}^*)}{n|\mathbf{H}|\tilde{f}(\mathbf{x}^*)}\int\{\tilde{K}(\mathbf{z}^*)\}^2 d\mathbf{z}^* + o\left(\frac{1}{n|\mathbf{H}|}\right). \tag{6}$$

Consequently, for the case where $\mathbf{H} = h\mathbf{I}_{D-1}$, the value of h which minimizes the asymptotic integrated version of $\mathcal{L}[\hat{\mathbf{m}}(\mathbf{x};\mathbf{H})]$ is

$$h_{\mathrm{opt}} = \left\{\frac{(D-1)\sum_{j=1}^{L-1} V_j}{4\sum_{j=1}^{L-1} U_j}\right\}^{\frac{1}{D+3}} n^{-\frac{1}{D+3}} \tag{7}$$

where V_j (U_j respectively) is the part of the leading term of the integrated variance (integrated squared bias resp.) of $\hat{m}_j^*$ not depending on h and n.

4 Real-simplicial regression

When the predictor, say X, is real-valued and the response takes values on $\mathcal{S}^D$, the real-simplicial regression function could be defined according to (3) by replacing $\mathbf{X} \in \mathcal{S}^D$ with $X \in \mathbb{R}$. Specifically, given the $\mathbb{R}\times\mathcal{S}^L$-valued random sample $(X_1,\mathbf{Y}_1)\ldots(X_n,\mathbf{Y}_n)$, we assume the model $\mathbf{Y}_i = \mathbf{m}(x)\oplus \boldsymbol{\epsilon}_i$, $i \in \{1,\ldots,n\}$, where the $\boldsymbol{\epsilon}_i$s, conditioned on the X_is, share the same properties discussed for the simplicial-simplicial regression case. Hence, a kernel estimator for $\mathbf{m}$ at $x \in \mathbb{R}$, could be defined as

$$\hat{\mathbf{m}}(x;h) := \mathrm{ilr}^{-1}\left(\hat{m}_1^*(x;h),\ldots,\hat{m}_{L-1}^*(x;h)\right)$$

with

$$\hat{m}_j^*(x;h) = \frac{\sum_{i=1}^n K_h(X_i - x)\mathrm{ilr}_j(\mathbf{Y}_i)}{\sum_{i=1}^n K_h(X_i - x)}, \quad j \in \{1,\ldots,L\}$$

where $K(\cdot)$ is a standard euclidean univariate kernel, and $h > 0$ is the bandwidth. Thus, denoting again the design density as f, and, letting, for $s \in \mathbb{Z}_+$

$$\gamma_s(K) := \int_{\mathbb{R}} K(u)u^s du, \quad \text{and} \quad \nu(K) := \int_{\mathbb{R}} K^2(u)du,$$

by applying standard results for bias and variance of kernel regression estimator (equipped with a second-order kernel and under classical conditions) to each $\hat{m}_j^*(x;h)$, $j \in \{1,\ldots,L\}$, we obtain

$$\mathcal{L}[\hat{\mathbf{m}}(x;h)] = \sum_{j=1}^{L-1}\left\{h^2\gamma_2(K)\left[\frac{m_j^{*\prime\prime}(x)}{2}+\frac{m_j^{*\prime}(x)f'(x)}{f(x)}\right]\right\}^2 + \sum_{j=1}^{L-1}\frac{\nu(K)\sigma_j^2(x)}{nhf(x)} + o\left(h^4+\frac{1}{nh}\right).$$

Thus, the value of h which minimizes the integrated version of the leading term of the above loss is

$$h_{\text{opt}} = \left\{\frac{\sum_{j=1}^{L-1} v_j}{4\sum_{j=1}^{L-1} u_j}\right\}^{\frac{1}{5}} n^{-\frac{1}{5}}$$

where v_j (u_j respectively) is the part of the leading term of the integrated variance (integrated squared bias resp.) of $\hat{m}_j^*$ not depending on h and n.

5 Application to real data

As an application, we use data from the "Kola Ecogeochemistry" Project, which is an environmental investigation in the Barents region whose main aim was to study pollution due to industrial activity and in which about 600 samples of soils were collected in four different layers: moss, O-horizon, B-horizon, C-horizon. See Reimann et al. (1998) for details. The whole data set is available in package StatDA of the statistical software R (see Filzmoser and Steiger (2009)).

We compare our nonparametric method with the model in Egozcue et al. (2012) where a regression from $\mathbb{R}^3$ to $\mathcal{S}^3$ is treated. Egozcue et al. (2012) considered three chemical elements: Fe (Iron), K (Potassium) and P (Phosphorus) taken from the O-horizon layer of soil. They studied the concentration of these three elements in dependence of latitude, longitude and elevation of the examined soil and found that elevation is significant in the regression model, while latitude is not significant and longitude is significant only at a level of $\alpha = 0.1$. For this reason we focus on the predictor elevation, say X.

In the ilr transformation, we use the same balances (coordinates) as Egozcue et al. (2012):

$$Y_1^* = \sqrt{\frac{2}{3}}\ln\frac{\text{Fe}}{\sqrt{\text{K}\cdot\text{P}}}, \quad Y_2^* = \frac{1}{\sqrt{2}}\ln\frac{\text{P}}{\text{K}}.$$

For the selection of the smoothing parameter, we choose h_j for $\hat{m}_j(x;h_j)$, $j \in \{1,2\}$, using the *leave-one-out cross validation* method, i.e. the values of h_js which minimize the function

$$CV_j(h_j) = \frac{1}{n}\sum_{i=1}^{n}(Y_{ji}^* - \hat{m}_{j(-i)}^*(X_i^*;h_j))^2$$

where Y_{ji}^* is the j-th component of $\mathbf{Y}_i$ and $\hat{m}_{j(-i)}^*(X_i^*;h_j)$ is the estimate at X_i, computed by leaving out the i-th observation $(X_i,\mathbf{Y}_i)$. For the case of the Epanechnikov kernel, we found $h_1 = 140.0$ and $h_2 = 73.5$.

For the parametric model, the results are

coordinate	parameter	estimated value	$t-$statistic	$p-$value
Y_1^*	intercept	0.5667	10.499	$< 2\cdot 10^{-16}$
Y_1^*	elevation	$5.227\cdot 10^{-4}$	2.177	0.0299
Y_2^*	intercept	-0.1532	-7.110	$3.23\cdot 10^{-12}$
Y_2^*	elevation	$6.640\cdot 10^{-4}$	6.928	$1.08\cdot 10^{-11}$

then, elevation is mildly significant for the first coordinate Y_1^* and highly significant for the second coordinate Y_2^*. Since Fe is supposed to be independent from elevation, P and K have to be responsible for the significance. In particular, the ratio P/K increases with the elevation.

In Fig. 1, linear and nonparametric estimates for m_1^* are represented, and they are quite similar.

Now, we focus on the second coordinate Y_2^*, the log-ratio of phosphorus over potassium. From Fig. 2 we can see that the fits are still similar for the first 350 m, but after this point their trends are opposite, and the kernel estimate decreases.

This somewhat surprising behavior has a sound interpretation as follows. We need to put our attention on an other variable, related with elevation: the vegetation on the ground. At the beginning we searched a relation between elevation and the lithology of the soil, but it doesn't seem to be any relation between them in the data, while there is one considering elevation and

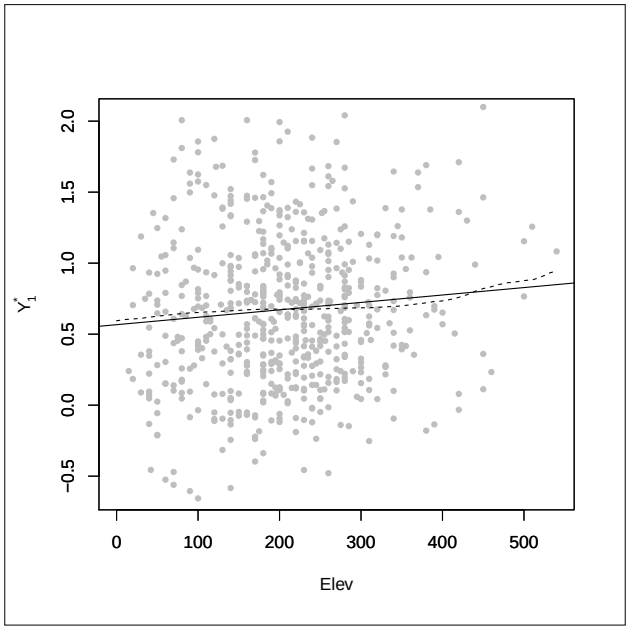


Figure 1: Estimates of regression of Y_1^* on elevation, obtained by parametric (solid line), and nonparametric (dashed line) methods.

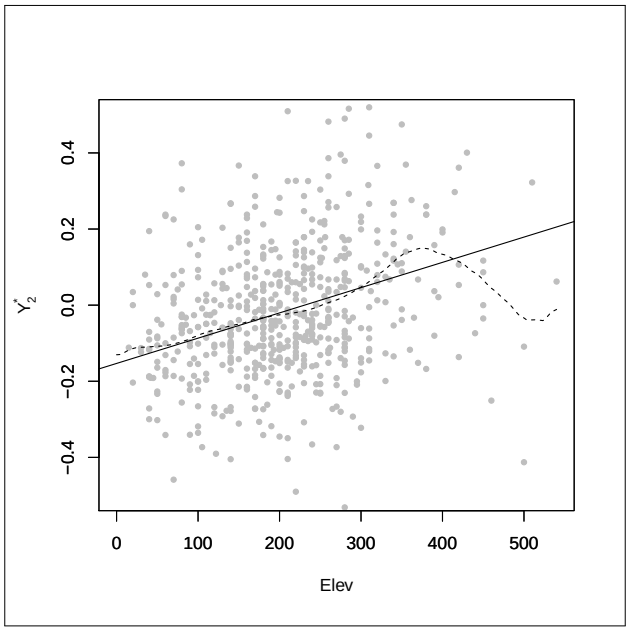


Figure 2: Estimates of regression of Y_2^* on elevation, obtained by parametric (solid line), and nonparametric (dashed line) methods.

ISBN: 978-3-200-03103-6

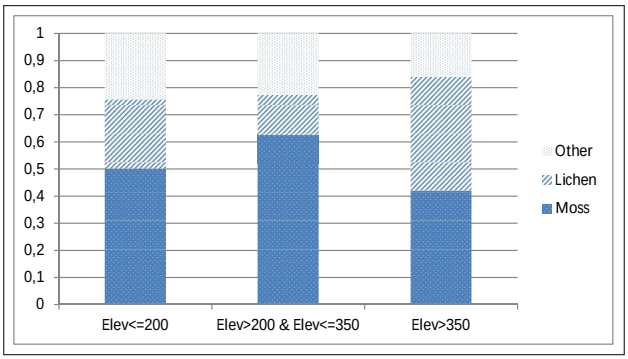


Figure 3: Vegetation on the ground with respect to three different levels of elevation

vegetation. See Fig. 3. The moss, which represents the majority of vegetation, first increases with elevation, but after 350 m, it has an inversion of the trend; the behavior of lichens is the opposite, at the beginning they decrease and then, after 350 m, they reach the same proportion as the moss. Since moss is a green plant while lichens are a symbiosis of fungus and algae, the first one requires more phosphorus over potassium. Phosphorus is a nutrient element for plants, potassium could be also a nutrient, but here it is present as a mineral. Consequently, the trend of the presence of moss and lichens with respect to elevation can explain the behavior of our coordinate $Y_2^* = \frac{1}{\sqrt{2}} \ln \frac{\mathrm{P}}{\mathrm{K}}$. The parametric model could explain only the main increasing trend, but not the behavior for high values of elevation, while our estimator seems to better capture the complexity of the phenomenon.

References

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society* *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).

Aitchison, J. and I. J. Lauder (1985). Kernel density estimation for compositional data. *Applied Statistics* *34*(2), 129–137.

Chacón, J.E., G. Mateu-Figueras and J. A. Martín-Fernández (2011). Gaussian kernels for density estimation with compositional data. *Computers & Geosciences* *37*(5), 702–711.

Egozcue, J.J., C. Barceló-Vidal, J.A. Martín Fernández, E. Jarauta-Bragulat, J.L. Díaz-Barrero and G. Mateu-Figueras (2011). Elements of simplicial linear algebra and geometry. In V. Pawlowsky-Glahn and A. Buccianti, editors, *Compositional Data Analysis. Theory and Applications.* John Wiley & Sons, Chichester (UK), p. 150.

Egozcue, J.J., J. Daunis-i-Estadella, V. Pawlowsky-Glahn, K. Hron and P. Filzmoser (2012). Simplicial regression. The normal model. *Journal of Applied Probability and Statistics* *6*(1 & 2), 87–108.

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.

Filzmoser, P. and B. Steiger (2009). *StatDA: Statistical Analysis for Environmental Data.* R package version 1.1.

Härdle, W. (1990). *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.

Hijazi,R.H. and R.W. Jernigan (2009). Modelling compositional data using Dirichlet regression models. *Journal of Applied Probability and Statistics* *6*(1 & 2), 87–108.

Mateu-Figueras, G., V. Pawlowsky-Glahn and C. Barceló-Vidal (2003). Distributions on the simplex. In *Compositional Data Analysis Workshop - CoDaWork'03.* http://ima.udg.es/Activitats/CoDaWork03/.

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.

Reimann, C., M. Äyräs, V. Chekushin, I. Bogatyrev, R. Boyd, P. d. Caritat, R. Dutter, T. Finne, J. Halleraker,O. Jæger, G. Kashulina, O. Lehto, H. Niskavaara, V. Pavlov, M. Räisänen, T. Strand, T. Volden (1998). Environmental geochemical atlas of the Central Barents Region. *Geological Survey of Norway (NGU), Geological Survey of Finland (GTK) and Central Kola Expedition (CKE)*. Special publication, Trondheim, Espoo, Monchegorsk.

Tolosana-Delgado,R. and K. G. Van Den Boogart (2011). Linear models with compositions in R. In V. Pawlowsky-Glahn and A. Buccianti, editors, *Compositional Data Analysis. Theory and Applications*. John Wiley & Sons, Chichester (UK), pp. 356–371.

Wand, M. P. and M. C. Jones (1995). *Kernel smoothing*. Chapman and Hall, London.

Objective comparison of mean with median and standard deviation with median absolute deviation for statistically contaminated samples of size 5-20 from Monte Carlo simulations and implications for data processing of three chemical elements in two international geochemical reference materials

SURENDRA P. VERMA[1], LORENA DÍAZ-GONZÁLEZ[2], and JOSÉ RAÚL GARCÍA-GILES[3]

[1] Departamento de Sistemas Energéticos, Instituto de Energías Renovables, Universidad Nacional Autónoma de México, Priv. Xochicalco s/no., Col. Centro, Apdo. 34, Temixco 62580, Mexico, spv@cie.unam.mx

[2] Departamento de Computación, Facultad de Ciencias, Universidad Autónoma de Estado de Morelos, Av. Universidad 1001, Chamilpa, Cuernavaca, Mor., 62209, Mexico.

[3] Licenciatura en Ciencias, Facultad de Ciencias, Universidad Autónoma de Estado de Morelos, Av. Universidad 1001, Chamilpa, Cuernavaca, Mor., 62209, Mexico

Monte Carlo simulations for sample sizes of five to twenty involving a simple statistical contamination of one datum show that both mean and standard deviation must be computed after the application of discordancy tests. Due to space limitations, only those simulations are reported in which the contaminant observation constituted the outlier to be tested for discordancy. The mean and standard deviation (outlier-based estimators) generally provide a better estimate of the population than the median and MAD (robust estimators), respectively. Reasons for the better performance of the outlier-based methods in comparison to the robust methods are suggested. A preliminary statistical procedure is also presented to estimate the central tendency and dispersion parameters for three elements (K, Th and U: the main heat-producing elements in the Earth) in two geochemical reference materials (Hawaiian basalt BHVO-1 from the U.S. Geological Survey and Japanese basalt JB-1 from the Geological Survey of Japan).

1 Introduction

In all experimental analytical or quantification work, it should be mandatory to estimate both central tendency and dispersion parameters (Verma, 2012). Two basically different approaches, known as "Outlier-based" and "Robust" methods, are available for estimating these statistical parameters (Barnett and Lewis, 1994; Miller and Miller, 2005; Verma, 2005; Maronna et al., 2006). In most cases probably due to the presence of discordant outlying observation(s), mean and median on one hand, and standard deviation and MAD on the other, do not agree with one another (e.g. Barnett and Lewis, 1994; Verma, 2005). This disagreement, therefore, requires an objective way to understand the relationship of these four statistical estimators with the central tendency and dispersion parameters of the sampled population.

It has been recommended that the mean and standard deviation parameters should always be computed after the application of discordancy tests for outliers (Barnett and Lewis, 1994; Verma, 1997, 2012; Verma and Díaz-González, 2012), which is generally not followed. Much simulation work remains to be done to better understand the importance of this recommendation. The present work involving Monte Carlo simulations is the first attempt in this direction.

Because the median and MAD belong to the category of robust methods, it is generally believed, without any precise Monte Carlo simulation evidence, that both of them provide unbiased estimates of central tendency and dispersion parameters, respectively, even in the presence of statistical contamination (Miller and Miller, 2005; Maronna et al., 2006). In fact, it is generally accepted that the robust parameters have larger breakdown points than the outlier-based parameters, e.g., both the median and MAD are considered to have the highest possible breakdown point of 50%, whereas the mean and standard deviation have the lowest breakdown point of only 0% (Maronna et al., 2006; Williams, 2011). However, from 1000 repetitions of Monte Carlo simulations, Carroll and Wegman (1975) inferred that none of the robust estimators are very robust over short-tailed distributions, but they also warned that more simulation work was needed. Other more recent simulation studies (e.g. Williams, 2011) with larger repetitions have shown that finite sample size dependent correction (multiplication) factors can be used to convert the MAD into the "sample" or "normal" (i.e., population) standard

deviation. Miller and Miller (2005) stated that MAD/0.6745 is a useful robust estimate of the population standard deviation (σ), but no evidence for such a statement was presented. Probably, this denominator value of 0.6745 came from the work of Hampel (1974) and corresponds to the limiting value for the sample size of infinity.

In the present work, we used Monte Carlo simulations to compare four statistical parameters – mean with median and standard deviation with MAD – in small samples of size 5-20 containing only one statistically contaminated datum on the upper side of the data array. We also present an example of compositional data for two international geochemical reference materials.

2 Monte Carlo procedure for constructing statistical samples

A well-tested simulation procedure by Verma and Quiroz-Ruiz (2006) was used for generating 101 streams of independently and identically distributed (IID) normal samples from the distribution centred at 0 and with standard deviation or variance of 1, i.e., from N(μ,σ) distribution, where population mean μ = 0 and population dispersion σ = 1. To construct statistically contaminated samples of size 5-20, two different independent streams were used, one being the original N(0,1) distribution and the other having the central tendency shifted by a parameter called δ, i.e., the contamination arose from the distribution N(δ,1). To generate random data from the distribution N(δ,1), we should add δ value to N(0,1); we will, therefore, call the new contaminant distribution as N(0+δ,1).

For samples of size n (5-20 in this work), one stream of N(0,1) was used to obtain n-1, i.e., 4-19 data and a different stream of N(0+δ,1) for the contamination by the remaining datum x^* to complete the array, i.e., to obtain the desired samples of size n, i.e., 5-20. Let us denote this combined initial array as $x_1, x_2, x_3, \ldots x_{n-2}, x_{n-1}, x_n$. For example, for constructing a contaminated sample of size five, i.e., n = 5, four data were drawn from a stream of N(0,1) and one from a shifted stream of N(0+δ,1). In other words, we contaminated only from the positive values of δ and added only one datum x^* from this location-shifted distribution N(0+δ,1) to the bulk of samples drawn from the original N(0,1) distribution. Thus, in this work our aim was to implement asymmetrical contamination on the upper side of the data array.

The (n-1) data sampled from N(0,1) distribution were first arranged from the lowest to the highest value, before adding the sole datum x^* from N(0+δ,1), which represented the most fundamental or the simplest asymmetrical statistical contamination. We denote this new combined ordered array as $x_{(1)}, x_{(2)}, x_{(3)}, \ldots x_{(n-2)}, x_{(n-1)}, x_{(n)}$. When the final statistically contaminated sample of size n was thus constructed to represent actual experiments, two distinct possibilities or events occurred as follows: (i) the contaminated datum $x^* = x_{(n)}$ was observed to be outside the ordered array on its upper side, i.e., x^* represented an upper outlier value $x_{(n)}$, an event called C type; and (ii) the contaminated datum $x^* \neq x_{(n)}$ was observed to be within the ordered array, i.e., x^* did not constitute an upper outlier value $x_{(n)}$, an event called $\hat{C}$ type.

Because of space limitations, the probabilities of generating these C type events from Monte Carlo simulations are not discussed in this paper; similarly, the results for $\hat{C}$ type events are not presented. These probabilities and their variations as a function of δ play an important role in understanding the results for zero contamination, which are also not presented here.

In our study, δ was varied from 0 to 10 as follows: 0(0.01)0.1(0.1)1(0.5)10. Therefore, for the comparison of the four statistical estimators evaluated in this study (mean, median, standard deviation, and MAD) could be considered for both uncontaminated (when δ = 0) and contaminated (when δ = 0.01-10) normal samples. However, we have left the discussion of uncontaminated samples for a more detailed paper. Unfortunately, the δ parameter is of academic interest only. Therefore, in our simulation experiments for each δ we computed a practical parameter called τ (Normalized distance of the outlier from the censored mean) proposed for the first time. For an ordered array $x_{(1)}, x_{(2)}, x_{(3)}, \ldots x_{(n-2)}, x_{(n-1)}, x_{(n)}$ of size n, the censored array was defined as the data array of size (n-1) as $x_{(1)}, x_{(2)}, x_{(3)}, \ldots x_{(n-2)}, x_{(n-1)}$, i.e., without considering the upper outlier or extreme

observation $x_{(n)}$. Mean ($\bar{x}_{(n-1)}$) and standard deviation ($s_{(n-1)}$) values of the censored array of size (n-1), i.e., for the array $x_{(1)}, x_{(2)}, x_{(3)}, \ldots x_{(n-2)}, x_{(n-1)}$ were then computed. The parameter τ (in fact, τ_C, because we are considering here only C type events) was calculated from equation 1 as follows.

$$\tau_C = (x_{(n)} - \bar{x}_{(n-1)})/s_{(n-1)} \qquad (1)$$

Because τ_C is defined in terms of the ordered data array, it is always possible to calculate it in any experiment, irrespective of the population N(μ,σ) from which the contaminant upper outlier $x_{(n)}$ was derived, i.e., we are not forced to know the theoretical parameter δ. Furthermore, it is a number, independent of the measurement units. Therefore, τ_C should result in a useful new parameter for the interpretation of actual data in all kinds of experiments.

3 Calculation of statistical parameters for simulated samples

Four parameters (mean, median, standard deviation, and MAD) were calculated for each statistical sample. Following Barnett and Lewis (1994) and Verma (2012), the term outlier is used here to refer to a datum that lies outside an ordered array of data, irrespective of whether or not it is interpreted as a discordant observation. The discordancy can be judged from the application of the multiple test method (MTM) originally proposed by Verma (1997), along with the new highly precise and accurate critical values (e.g., Verma et al., 2008). When an outlier or extreme observation is detected as discordant, it is called a discordant outlier (Verma, 2012).

Appropriate discordancy tests for an upper outlier of the single-outlier type (N1U, N2, N4U, N7U, N8, N9U, N10U, N14, and N15; Barnett and Lewis, 1994; Verma et al., 2009) were applied at the strict 99% confidence level to each data array and the mean and standard deviation values were calculated after this application. Two distinct possibilities arose as follows: (i) the outlier was not detected as discordant (called $\overline{D}$ type), and (ii) it was detected as discordant (called D type).

On the other hand, because both median and MAD pertain to the category of robust methods, these statistical parameters were calculated from the initial datasets without the application of the discordancy tests. Similarly, for completeness and for understanding the importance of discordancy tests, both mean and standard deviation were also calculated from the initial data arrays, although they should always be calculated after the application of discordancy tests.

The replication was set at 10,000,000 for each simulation experiment. The whole process was repeated 20 to 101 times using totally different N(0,1) and N(0+δ,1) streams. These sets of results were used to estimate the final uncertainty (99% confidence intervals of the mean), which were extremely small, demonstrating very high precision of our Monte Carlo simulations. The accuracy of our simulations was also similarly high as already demonstrated by Verma and Quiroz-Ruiz (2006).

4 Comparison of statistical parameters

4.1 Evaluation of statistical parameters in terms of τ_C

Figures 1 and 2 show the behaviour of the central tendency and dispersion estimators, respectively, as a function of τ_C. In our simulation experiments, the expected values of these two parameters were 0 and 1, respectively.

4.1.1 Evaluation of the mean and median in terms of τ_C

When the discordancy tests were not applied to the data array, for C type events (Figure 1a-d) the mean $\bar{x}_C$ increased very rapidly from the expected value of 0, reaching about 0.5 at τ_C = 3.8 (n = 5), 5.6 (n = 10), 8.0 (n =

15), and 10.4 ($n = 20$). In fact, for larger values of τ_C, the mean $\bar{x}_C$ continued to depart from the expected value of zero. This clearly implies that discordancy tests must always be applied before the calculation of the mean, being an outlier-based parameter (Verma, 1997, 2012).

Nevertheless, depending on the values of τ_C and efficiency of discordancy tests (for the evaluation of efficiency see Barnett and Lewis, 1994), their application may result in either the tests do not detect the upper observation as discordant ($\overline{D}$ type result) or they do so (D type result). The sample mean $\bar{x}_{C\overline{D}}$ values for $n = 5$, 10, 15, and 20 (Figure 1a-d) varied from about 0.0-0.3 for τ_C = 1.8-2.1, 2.0-2.8, 2.1-3.2, and 2.2-3.4, respectively. The median $\tilde{x}_{C\overline{D}}$ also similarly departed from the true value of 0 for $n = 5$, 10, 15, and 20 (Figure 1a-d), because it varied from about 0.0-0.3 for τ_C = 1.8-2.1, 2.0-3.5, 2.1-3.7, and 2.2-3.8, respectively.

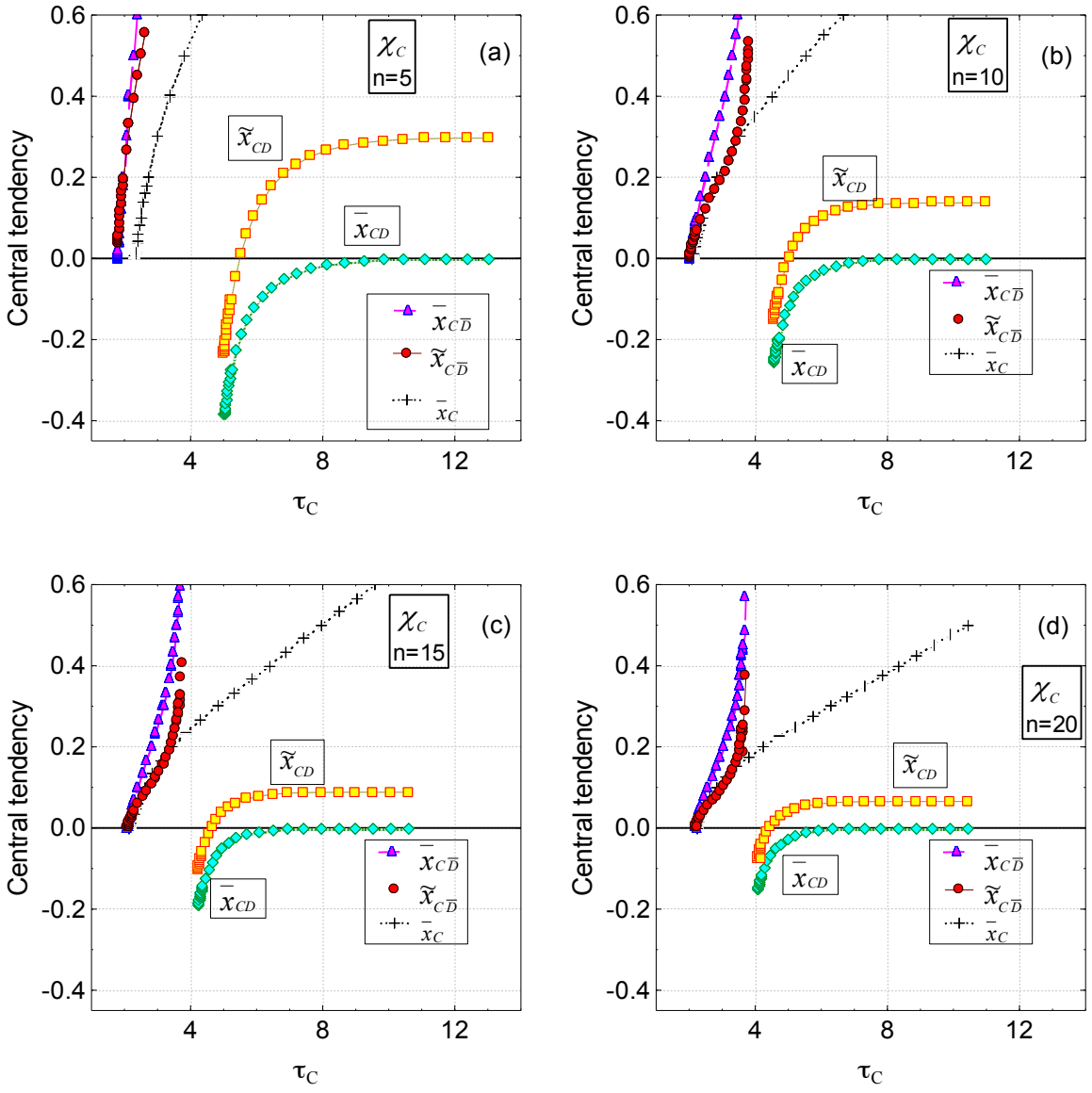


Figure 1. Central tendency or location parameters for C type events in which the contaminant x^* was used in the test statistics, as a function of τ_C. The solid horizontal line at zero value of the y-axis is the "true" population central tendency or the expected value. (a) sample size five ($n = 5$); (b) sample size ten ($n = 10$); (c) sample size fifteen ($n = 15$); and (d) sample size twenty ($n = 20$).

For larger values of τ_C both mean $\bar{x}_{C\bar{D}}$ and median $\tilde{x}_{C\bar{D}}$ continued to depart from the true central tendency until τ_C reached the values appropriate for the discordancy tests to be "successful" in detecting the upper outlier as discordant (unpublished simulation results on the efficiency of tests by our group); these parameters are referred to as $\bar{x}_{CD}$ and $\tilde{x}_{CD}$ (Figure 1a-d). Discordancy tests became successful in detecting the upper outlier $x_{(n)}$ (contaminant x^*) as discordant for τ_C > 5.0, 4.6, 4.3, and 4.1, for n = 5, 10, 15, and 20, respectively.

For n = 5 and contaminated samples, the mean $\bar{x}_{CD}$ varied from about -0.38 to -0.01 for τ_C of about 5.0 to 9.4, but became practically the expected value of 0.00 for τ_C > 9.4 (Figure 1a). The median ($\tilde{x}_{CD}$), however, rapidly changed from -0.23 for τ_C = 5.0 to +0.29 for τ_C = 9.8 and kept on increasing (i.e., departed from the true value of zero), although very slowly, to still somewhat higher values. In other words, the median never stabilized at the expected value of 0.00 (Figure 1a). This is an unexpected result, contrary to the general belief about the robustness of the median (Miller and Miller, 2005; Maronna et al., 2006). We infer a better behaviour of the mean $\bar{x}_{CD}$ as compared to the median $\tilde{x}_{CD}$ as judged from the closeness of $\bar{x}_{CD}$ to the true value of 0 for τ_C > 7.0 (Figure 1a).

Our results are actually true for other sample sizes as well (Figure 1b-d). For n = 10, 15, and 20, respectively, the mean $\bar{x}_{CD}$ values varied from -0.25 to -0.01, -0.19 to -0.01, and -0.15 to -0.01 for τ_C = 4.6-6.8, 4.2-6.1, and 4.1-5.5. $\bar{x}_{CD}$ values were practically indistinguishable from 0.00 for higher values of τ_C, whereas for the same n and τ_C, the median $\tilde{x}_{CD}$ values varied from -0.15 to +0.14, -0.08 to +0.09, and -0.08 to +0.07 (Figure 1b-d).

Therefore, an important result of our Monte Carlo simulations is that mean $\bar{x}_{CD}$ would be a more reliable estimate of the population central tendency in more cases that the median $\tilde{x}_{CD}$. Thus, for n = 5 and τ_C > 6.2, n = 10 and τ_C > 5.8, n = 15 and τ_C > 5.4, and n = 20 and τ_C > 5.2, respectively, when the discordancy tests do detect a contaminant observation as discordant, the mean ($\bar{x}_{CD}$) would be closer to the population central tendency value than the median ($\tilde{x}_{CD}$). However, for contaminated samples with n = 5, 10, 15, and 20, and for some what narrow range of τ_C = 5.0-5.9, 4.6-5.3, 4.2-4.9, and 4.1-4.8, respectively, $\tilde{x}_{CD}$ would provide an estimate of the central tendency, which is probably still biased but is better than $\bar{x}_{CD}$.

4.1.2 Evaluation of standard deviation and median absolute deviation in terms of τ_C

When the discordancy tests were not applied to the data array, the standard deviation s_C for C type events and n = 5, 10, 15, and 20 (Figure 2b-d), also departed from the expected value of 1, but varied from 0.94-4.6, 0.97-3.3, 0.98-2.8, and 0.99-2.4 for τ_C = 2.3-13.8, 2.1-11.1, 2.2-10.6, and 2.2-10.4, respectively (for the sake of legibility, some of these results are not actually presented in Figure 2). This, once again, implies that discordancy tests must always be applied before the calculation of this outlier-based parameter as suggested by Verma (1997, 2012).

The sample standard deviation $s_{C\bar{D}}$ values (after an "unsuccessful" application of discordancy tests) for n = 5, 10, 15, and 20 (Figure 2a-d) varied from about 0.94-1.2, 0.97-1.2, 0.98-1.2, and 0.99-1.2 for τ_C = 1.8-2.1, 2.0-2.5, 2.1-2.8, and 2.2-2.9, respectively. The median absolution deviation $MAD_{C\bar{D}}$ departed even more from the true value of 1 for n = 5, 10, 15, and 20 (Figure 2a-d), because it varied from about 0.59-1.2, 0.62-1.2, 0.64-1.2, and 0.65-1.1 for τ_C = 1.8-2.1, 2.0-2.5, 2.1-2.8, and 2.2-2.9, respectively. Fortuitously, however in Figure 2a-d, both $s_{C\bar{D}}$ and $MAD_{C\bar{D}}$ were close to the true value of 1 for certain values of τ_C. Thus, for n = 5, 10, 15, and 20, we observed that $s_{C\bar{D}}$= 1.0 for τ_C = 1.9, 2.1, 2.2, and 2.2, respectively, whereas $MAD_{C\bar{D}}$ did so for τ_C = 2.5, 3.5, 3.6, and 3.7, respectively.

When the discordancy tests successfully detected the upper outlier $x_{(n)}$ (contaminant x^*) as discordant, both s_{CD} and MAD_{CD} departed from the true dispersion parameter value of 1 (Figure 2a-d). However, s_{CD}

departed significantly less than MAD_{CD}. For contaminated samples of n = 5, s_{CD} and MAD_{CD} varied, respectively, from 0.42-0.92 and 0.35-0.81, for τ_C = 5.0-13.0 (Figure 2a). Note that the final sample size for this kind of samples was four, instead of five, because the contaminant x^* was identified as discordant. For larger values of n (Figure 2b-d), s_{CD} was always closer to the expected value of 1 as compared to MAD_{CD}. For initial sample size of n = 10, 15, and 20, s_{CD} varied from 0.57-0.97, 0.68-0.98, and 0.74-0.99 for τ_C = 4.6-11.0, 4.3-10.6, and 4.1-10.4, respectively, whereas MAD_{CD} varied from 0.39-0.71, 0.47-0.70, and 0.52-0.69, respectively, for the same values of τ_C as for s_{CD}.

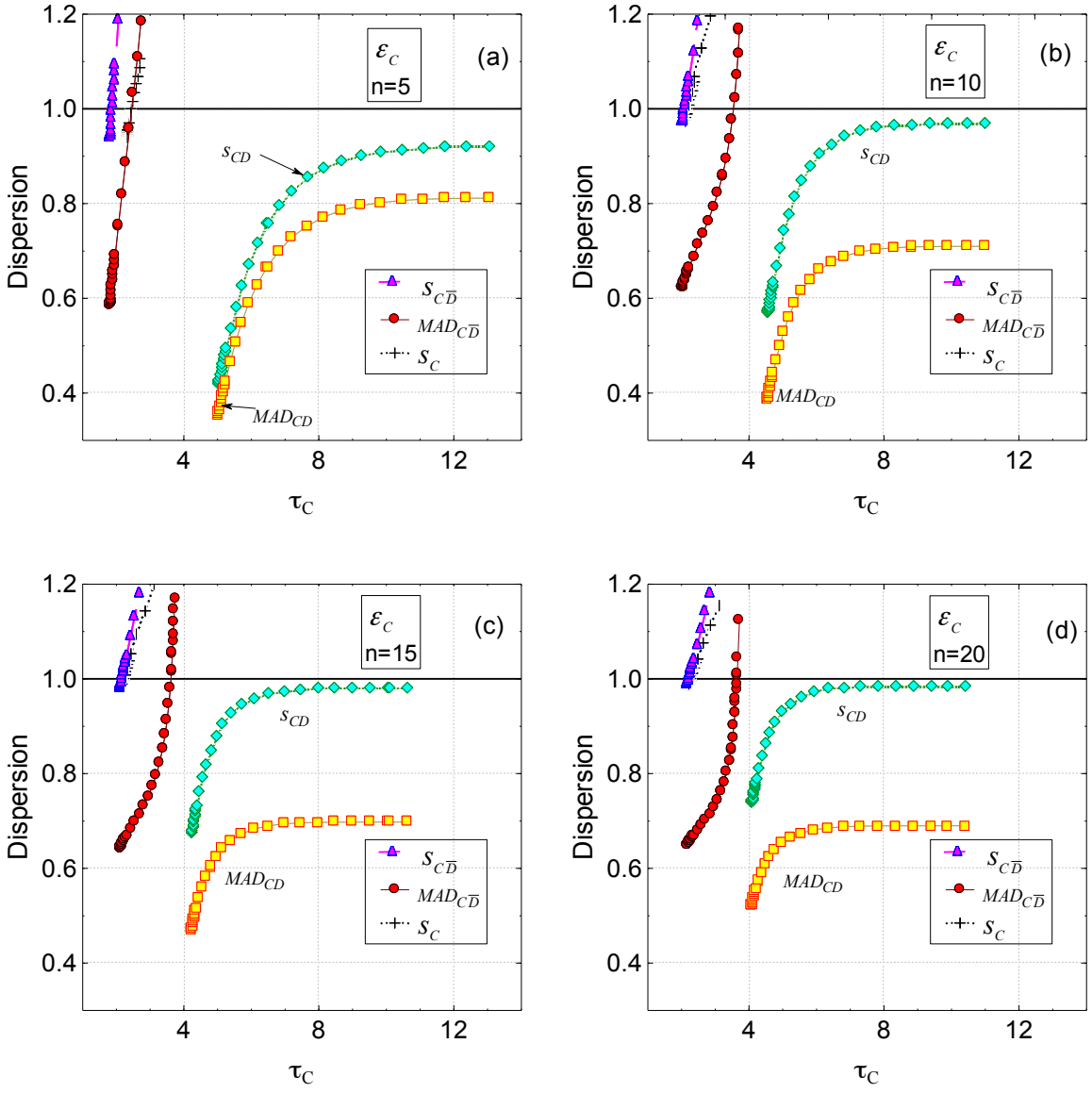


Figure 2. Dispersion or scale parameters for C type events in which the contaminant x^* was used in the test statistics, as a function of τ_C. The solid horizontal line at the value of one of the y-axis is the "true" population standard deviation or the expected value of the dispersion parameter. (a) sample size five (n = 5); (b) sample size ten (n = 10); (c) sample size fifteen (n = 15); and (d) sample size twenty (n = 20).

Because the sample standard deviation s_{CD} varied in a complex way (Figure 2a-d), it cannot be easily converted to the population dispersion ($\sigma = 1$) by any simple correction. Nevertheless, it is closer to the expected value of one for larger sample sizes (compare, for example, the results for $n = 5$ with $n = 20$; Figure 2a-d).

Another important result of our Monte Carlo simulation modelling concerns the median absolution deviation MAD_{CD} in that no simple relationship, such as that proposed by Hampel (1974), Miller and Miller (2005), or Williams (2011), can correct it to reach the population dispersion estimate. Instead, complex τ and n dependent correction factors or equations should be proposed, in future, to correct both s_{CD} and MAD_{CD}. Thus for $n = 5$, the expected population dispersion σ can be obtained from MAD_{CD} by multiplying it by about 2.836 for τ_C=5.0 to near asymptotic multiplication factor value of about 1.232 for τ_C=13.0 (Figure 2a). Similarly, for $n = 10$, 15, and 20 the correction factors for MAD_{CD} would be, respectively, about 2.583 for τ_C=4.6 to 1.409 for τ_C=11.0 (Figure 2b), about 2.125 for τ_C=4.2 to 1.433 for τ_C=10.6 (Figure 2c), and about 1.908 for τ_C=4.1 to 1.450 for τ_C=10.4 (Figure 2d).

5 Interpretation of simulation results

We have evaluated the addition of only one contaminant observation arising from a location-shifted distribution N(0+δ,1) that is added to the array of (n-1) data drawn from the distribution N(0,1), where δ was assumed to be positive (up to 10). Very similar considerations will apply if the contamination were assumed to originate from a scale-shifted distribution such as N(0, εσ) or both location- and scale-shifted distribution N(0+δ,εσ) where $\sigma = 1$. The same arguments will apply for the general distribution N(μ,σ) when the contamination arises from N(μ+δμ,εσ), where the shift or contaminant parameter δμ could be positive or negative, μ any positive or negative value, σ any positive value, and $\varepsilon < 1$ or > 1.

5.1 A possible explanation for the poor performance of robust estimators

The poor performance of both median ($\tilde{x}_{CD}$ and $\tilde{x}_{C\bar{D}}$) and median absolute deviation (MAD_{CD} and $MAD_{C\bar{D}}$) is well documented in this work and contradicts the general belief about their "good" (robust) performance (Miller and Miller, 2005; Maronna et al., 2006). It can probably be understood in terms of what an asymmetrical statistical contamination does to the statistical samples. For robust estimators, it is the position of the data in the ordered array that is important rather than most of the data values. The fact, as explained below, is that an asymmetrical contamination causes a shift of the location or position of the median or MAD towards the location of the datum or data resulting from the contamination.

We will highlight our arguments about the performance of median and MAD from the following considerations. For an uncontaminated ordered array of n-1 data, the median is represented by the $[\{(n\text{-}1)+1\}/2]^{th}$ data value for odd values of (n-1) or by the average of $[(n\text{-}1)/2]^{th}$ and $[\{(n\text{-}1)/2\}+1]^{th}$ data values for even values of (n-1). When this array is contaminated, for example, by one datum drawn from a different (e.g., location-shifted) population, the initial array of size (n-1) would become of size n. The median is then given, respectively, by the average of $[n/2]^{th}$ and $[\{n/2\}+1]^{th}$ data values for odd values of (n-1), i.e., for even values of n, and by $[\{n+1\}/2]^{th}$ data value for even values of (n-1), i.e., for odd values of n.

As an example of the sample size of five when four data are drawn from the dominant uncontaminated population and one datum from the contaminated population, we will have to calculate the median as the third value in the ordered array of these five data. However, because the corresponding uncontaminated sample was of size four only, the median for this uncontaminated sample will be given by the average of the second and third data.

Therefore, any kind of asymmetric contamination would cause a shift in the position of the median value, which would result in a biased estimate of this robust parameter. The same arguments are valid for the biased behaviour of the MAD.

That the size of the actual experimental (statistical) sample, in fact, grows as a result of the statistical contamination is certainly the best and correct way to interpret such data arrays, instead of supposing that some

experimental observations are replaced by contaminated datum or data. The erroneous "replacement" mechanism envisioned by some people is probably responsible for their vision of robust behaviour of the median and MAD parameters.

5.2 A misconception about the outlier-based methods

Another common problem is related to the false argument sometimes put forth against the use of discordant outlier-based methods because, according to this belief, the discordancy tests may eliminate the only valid observation(s) as discordant. We must remember that discordant observations and not legitimate ones occur at either extreme of an ordered univariate data array (Barnett and Lewis, 1994). Valid observations, especially those which are close to the central tendency value, are likely to occupy inner positions of an ordered array of data, irrespective of whether it is an uncontaminated or a contaminated sample, i.e., a strictly normally distributed data array in which all observations were drawn from a single distribution $N(\mu,\sigma)$, or a data array in which a large number of observations came from $N(\mu,\sigma)$ and a small number originated from a shifted distribution $N(\mu+\delta\mu,\sigma)$, $N(\mu,\varepsilon\sigma)$, or $N(\mu+\delta\mu,\varepsilon\sigma)$. The observations of inner positions are never tested for discordancy. Therefore, valid observations are always likely to be retained in any application of discordancy tests.

In this context, we may ask the scientific community, especially to those who prefer robust methods in comparison to outlier-based methods, the following question: “Are there any statistical methods in the category of robust methods that identify outlying observations as legitimate?” Obviously, there are no such methods. For example, the median is based on just one or two values located at the central position of odd or even numbered ordered data arrays, respectively; it does not take into account any of the outer” values other than the one or two "inner" values. All other values, including the outliers, are always ignored. The same is true of all other robust central tendency estimators, such as mean quartile, trimean, and Gastwirth´s mean; for other estimators such as trimmed mean or Winsorised mean there is no objecting way to decide about the extent of trimming or Winsorisation (Barnett and Lewis, 1994; Verma, 2005).

We further emphasize that the statistically correct way to use the mean and standard deviation as indicators of the population parameters is to evaluate the data array from discordancy tests (Verma and Díaz-González, 2012), as confirmed in this work from Monte Carlo simulations. Sometimes, it is decided "a priori" in favour of the robust methods that such an application of outlier-based methods is impractical for some applications (Williams, 2011), it is certainly not so now with the availability of the new software DODESSYS (Verma and Díaz-González, 2012).

5.3 Future work involving Monte Carlo simulation modelling

For two or more upper or lower outliers, or more complex contamination involving both inner and outer contaminants in samples of sizes five to twenty as well as in larger samples, additional simulation work is required before the entire topic of an objective comparison of different statistical parameters may be considered complete. Other kinds of more complex contaminant distributions, such as $N(\mu+\delta\mu,\varepsilon\sigma)$, should also be investigated to better understand the τ_C and $\tau_{\hat{C}}$ parameters in terms of theoretical concepts of both δ and ε.

Although the recent software DODESSYS proved to be useful in the present work, still newer software UDASYS (Univariate Data Analysis SYStem; Verma et al., in preparation), which allows a highly efficient use of both robust and outlier-based methods, as well as of discordancy and significance tests, would be an asset to routinely and efficiently apply the concepts put forth in the present work.

6 Compositional data for geochemical reference materials

We are quite far from fully understanding the implications of the results of our Monte Carlo simulations for compositional data. The discordancy tests should strictly be applied to log-transformed data and not to crude compositions (Verma, 2012). However, getting back to compositions as we, the geochemists, understand them is a difficult task. Nevertheless, we selected three chemical elements U, Th, and K, which are important for Earth´s

evolution models, because these elements constitute the main heat producing elements. Instead of evaluating log-transformed data (Aitchison, 1986; Egozcue et al., 2003), such data have been traditionally processed as crude univariate compositional data (e.g., Gladney et al., 1991; Govindaraju, 1994; Imai *et al*., 1996; Verma, 1997; Velasco-Tapia et al., 2001; Marroquín-Guerra et al., 2009; Pandarinath, 2009).

Log-ratio transformation has been used for compositional data by several workers (e.g., Pawlowsky-Glahn and Egozcue, 2006; Buccianti et al., 2006; Filzmoser et al., 2009), but the concept of discordant outliers has not been yet extensively discussed in such log-transformed data, except in multi-dimensional diagrams for igneous rocks (e.g., Verma et al., 2013 and references therein).

The data for these elements in Hawaiian basalt BHVO-1 (issued by the U.S. Geological Survey) were compiled from Kelley et al. (2003), Lee et al. (2007), Makishima and Nakamura (2006), Manikyamba et al. (2005), Price et al. (1997), Rolland et al. (2002), and the internet site www.georem.mpch.mainz.gwdg.de. Similarly, the data used for Japanese basalt JB-1 (issued by the Geological Survey of Japan) were compiled from Makishima and Nakamura (2006) and the internet site www.aist.go.jp/RIODB/geostand. The individual data (only the complete data for the three elements U, Th and K, n=8 for BHVO-1 and n=10 for JB-1) and their log-transformed ratios (additive log-ratio, alr by Aitchison, 1986 and isometric log-ratio, ilr by Egozcue et al., 2003) are summarised in Table 1.

RM	Heat-producing elements ($\mu g.g^{-1}$)			Additive log-ratio (no units)		Isometric log-ratio (no units)		Reference
	K	Th	U	alr1	alr2	ilr1	ilr2	
BHVO-1	4317	1.247	0.4083	9.2660	1.1165	5.7626	4.2386	Kelley et al. (2003)
	4267	1.23	0.41	9.2503	1.0986	5.7641	4.2249	Lee et al. (2007)
	4358	1.28	0.429	9.2261	1.0932	5.7509	4.2128	Lee et al. (2007)
	4616	1.23	0.428	9.2858	1.0556	5.8196	4.2219	Makishima and Nakamura (2006)
	4151	1.08	0.41	9.2226	0.9686	5.8365	4.1605	Manikyamba et al. (2005)
	4317	1.00	0.40	9.2866	0.9163	5.9187	4.1653	Price et al. (1997)
	9049§	1.10	0.41	10.0020§	0.9869	6.3746§	4.4862§	Rolland et al. (2002)
	4400	1.23	0.409	9.2834	1.1011	5.7858	4.2394	www.georem.mpch.mainz.gwdg.de
$(\bar{x}_{CD})_{BHVO1}$	4347§	1.175	0.4130	9.260§	1.042	5.805§	4.209§	This work
$(\tilde{x}_{CD})_{BHVO1}$	4338	1.23	0.41	9.275	1.074	5.803	4.223	This work
$(s_{CD})_{BHVO1}$	143	0.100	0.0101	…	…	…	…	This work
$(MAD_{CD})_{BHVO1}$	66	0.0355	0.00135	…	…	…	…	This work
JB-1	11954	9.19	1.65	8.8881	1.7173	5.0705	4.3296	Makishima and Nakamura (2006)
	10377	9.4	1.70	8.7167	1.7101	4.9544	4.2567	www.aist.go.jp/RIODB/geostand
	10800	8.9	1.60	8.8173	1.7160	5.0213	4.3002	www.aist.go.jp/RIODB/geostand
	10875	8.9	1.60	8.8242	1.7160	5.0262	4.3030	www.aist.go.jp/RIODB/geostand
	11124	9.04	1.80	8.7291	1.6139	5.0312	4.2225	www.aist.go.jp/RIODB/geostand
	11207	8.1	1.70	8.7937	1.5612	5.1141	4.2274	www.aist.go.jp/RIODB/geostand
	11207	8.8	2.00	8.6311	1.4816	5.0555	4.1285	www.aist.go.jp/RIODB/geostand
	11700	9.0	1.77	8.7964	1.6262	5.0700	4.2550	www.aist.go.jp/RIODB/geostand
	11788	7.8	0.32§	10.5143§	3.1936§	5.1765	5.5962§	www.aist.go.jp/RIODB/geostand
	12120	8.56	1.40	9.0662	1.8106	5.1304	4.4404	www.aist.go.jp/RIODB/geostand
$(\bar{x}_{CD})_{JB1}$	11315	8.77	1.691	8.807	1.661	5.065	4.274	This work
$(\tilde{x}_{CD})_{JB1}$	11207	8.9	1.70	8.796	1.710	5.063	4.257	This work
$(s_{CD})_{JB1}$	561	0.49	0.165	…	…	…	…	This work
$(MAD_{CD})_{JB1}$	450	0.215	0.085	…	…	…	…	This work

Table 1: Crude and transformed geochemical data and preliminary basic statistics for two reference materials (§ one discordant outlier was observed from single-outlier type tests at 99% confidence level; the corresponding τ_C were for BHVO-1: 33.0, 26.8, 9.6, and 8.4; for JB-1: -8.3, 14.0, 15.4, and 15.4; "…" implies that these parameters are not recommended to be calculated, see Filzmoser et al., 2009). All symbols are explained in the text.

The data were processed in DODESSYS software for identifying discordant observations in crude as well as log-transformed variables. Single-outlier type discordancy tests were applied at the strict 99% confidence level (Verma and Díaz-González, 2012). These results are also summarised by indicating the discordant observation by the symbol "§" in Table 1. The mean values marked by § in the lower part for each reference material of Table 1 may be better estimates of the central tendency parameter. Mean and standard deviation values without the application of discordancy tests are not presented in Table 1. Note that standard deviation for all three elements in both samples is higher than MAD (Table 1).

We also processed the inversion of the mean and median values for the alr and ilr statistics of Table 1. Similarly, the ratios were also calculated directly from the crude compositional data and their mean values were estimated after the application of discordancy tests, whereas their median was calculated from the complete dataset. The results are summarised in Table 2. The mean and median values calculated from the inversion differed from the corresponding values obtained from the ratio data.

The crude mean and median values (Table 1) were also used for calculating the ratios of Table 2; these values are included in brackets. Note that for ratios these statistical parameters were exactly the same as those obtained from the inversion of the log-ratio transformations.

RM	Heat-producing elements (no units)				Inverted ratios from alr (no units)		Inverted ratios from ilr (no units)		Reference
	K/U	Th/U	K/Th	$(ThK)^{0.5}/U$	K/U	Th/U	K/Th	$(ThK)^{0.5}/U$	
BHVO-1									
$\bar{x}_{CD}$	10514 (10525)	2.84 (2.85)	3689 (3700)	173 (173)	10525	2.85	3700	173	This work
$\tilde{x}_{CD}$	10573 (10580)	2.93 (3.0)	3577 (3527)	176 (178)	10580	3.0	3527	178	This work
JB-1									
$\bar{x}_{CD}$	6727 (6691)	5.29 (5.19)	1295 (1290)	189 (186)	6691	5.19	1290	186	This work
$\tilde{x}_{CD}$	6680 (6592)	5.55 (5.24)	1287 (1259)	189 (186)	6592	5.24	1259	186	This work

Table 2. Central tendency of element ratios obtained from handling of crude compositions as well as log-transformed procedures for two reference materials.

We have applied single-outlier type discordancy tests to our data, treating them as univariate samples. Nevertheless, Barnett and Lewis (1994) and Rencher (2002) provided details on the Wilks´ statistic (Wilks, 1963) for detecting single or more outliers in a multivariate normal distribution. They also indicated that the critical values by Wilks (1963) are only approximate. Jennings and Young (1988) simulated more precise critical values for one or more outliers in such multivariate distributions. Rencher (2002) also presented the equations to convert the Wilks´ statistic to the F test statistic. We plan to apply in future suitable tests for multivariate normal distributions.

7 Conclusions

The main conclusions of our Monte Carlo simulations of asymmetrical statistical contamination can be summarised as follows:

(1) The outlier-based methods require the application of discordancy tests prior to the calculation of mean and standard deviation. The common practice of simply calculating these parameters from a series of data must therefore be avoided.

(2) For C type events, the probability of detecting an upper outlier (a contaminant observation x^*) as discordant increases very rapidly with increase of τ_C as well as of sample size (5-20); this probability achieves the highest value of one. This is a very interesting inference fully consistent with our wishful thinking of detecting this contaminant outlying observation as discordant.

(3) The robust parameters median and MAD generally provide biased estimates. This conclusion is contrary to the general belief about an overall good performance of these robust parameters.

(4) The outlier method based mean and standard deviation, particularly after a successful application of discordancy tests, are generally preferable to the robust estimators median and MAD.

(5) Additional simulation work is required to cover more complex types of asymmetrical contamination and to propose new ways to interpret laboratory data. This will provide further guidelines for the correct handling of compositional data.

Acknowledgements

The participation of Raúl García-Giles was made possible through a scholarship as "Ayudante de Investigador" of the first author (SPV) from the Sistema Nacional de Investigadores. We are grateful to Alfredo Quiroz Ruiz for guiding us in the simultaneous use of eight processors for computations in the initial stage of this research and to Rene Cruz-Huicochea to participate in frequent discussions and streamline the development of new software UDASYS that will be used in our future work on these lines. This work was partly supported by project IN104813 (UNAM) to the first author.

References

Aitchison, J. (1986). The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK), 416 p.

Barnett, V. and T. Lewis (1994). Outliers in Statistical Data (3rd edition). Wiley, Chichester (UK), 584 p.

Buccianti A., G. Mateu-Figueras, V. Pawlowsky-Glahn (2006). Frequency distribution and natural laws in geochemistry. A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn, editors, Compositional data analysis in the Geosciences: from theory to practice. The Geological Society of London Special Publication 264, London (UK), pp. 175-189.

Carroll, R. J. and E. J. Wegman (1975). A Monte Carlo study of robust estimators of location. Institute of Statistics Mimeo Series #1040, 45 p.

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.

Gladney, E. S., E. A. Jones, E. J. Nickell, and I. Roelandts (1991). 1988 compilation of elemental concentration data for USGS DTS-1, G-1, PCC-1, and W-1. Geostandards Newsletter 15 (2), 199-396.

Govindaraju, K. (1994). 1994 Compilation of working values and sample description for 383 geostandards. Geostandards Newsletter 18 (Special Issue), 1-158.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. Journal of Statistical Computation and Simulation. 69 (346), 383-393.

Filzmoser P., K. Hron, and C. Reimann (2009). Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Science of the Total Environment 407, 6100-6108.

Imai, N., H. Sakuramachi, S. Terashima, S. Itoh, and Ando, A. (1996). 1996 Compilation of analytical data on nine GSJ geochemical reference samples, "sedimentary rock series". Geostandards Newsletter 20 (2), 165-216.

Jennings, L. W. and D. M. Young (1988). Extended critical values of the multivariate extreme deviate test for detecting a single spurious observation. Communications in Statistics - Simulation and Computation 17 (4), 1359-1373.

Kelley, K. A., T. Plank, J. Ludden, and H. Staudigel (2003). Composition of altered oceanic crust at ODP sites 801 and 1149. Geochemistry Geophysics Geosystems 4, 1-21.

Lee, C.-T. A., Q.-Z. Ying, A. Lenardic, A. Agranier, C. J. O'Neill, and N. Thiagarajan (2007). Trace-element composition of Fe-rich residual liquids formed by fractional crystallization: Implications for the Hadean magma ocean. Geochimica et Cosmochimica Acta 71 (14), p. 3601-3615.

Makishima, A. and E. Nakamura (2006). Determination of major, minor and trace elements in silicate samples by ICP-QMS and ICP-SFMS applying isotope dilution-internal standardisation. Geostandards and Geoanalytical Research 30 (3), 245-271.

Manikyamba, C., S. M. Naqvi, D. V. S. Rao, M. R. Mohan, T. C. Khanna, T. G. Rao, and G. L. N. Reddy (2005). Boninites from the Neoarchaean Gadwal greenstone belt, eastern Dharwar craton, India: implications for Archaean subduction processes. Earth and Planetary Science Letters 230 (1-2), 65-83.

Maronna, R. A., R. D. Martin, and V. J. Yohai (2006). Robust Statistics Theory and Methods. Wiley, Chichester (UK), 403 p.

Marroquín-Guerra, S. G., F. Velasco-Tapia, and L. Díaz-González (2009). Evaluación estadística de Materiales de Referencia Geoquímica del Centre de Recherches Pétrographiques et Géochimiques (Francia) aplicando un esquema de detección y eliminación de valores desviados. Revista Mexicana de Ciencias Geológicas 26 (2), 530-542.

Miller, J. N. and J. C. Miller (2005). Statistics and Chemometrics for Analytical Chemistry (3rd edition). Pearson Prentice Hall, Essex (UK), 271 p.

Pandarinath, K. (2009). Evaluation of geochemical sedimentary reference materials of the Geological Survey of Japan (GSJ) by an objective outlier rejection statistical method. Revista Mexicana de Ciencias Geológicas 26 (3), 638-646.

Pawlowsky-Glahn, V. and J. J. Egozcue (2006). Compositional data and their analysis: an introduction. In A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn, editors, Compositional data analysis in the Geosciences: from theory to practice. The Geological Society of London Special Publication 264, London (UK), pp. 1-10.

Price, R. C., C. M. Gray, and F. A. Frey (1997). Strontium isotopic and trace element heterogeneity in the plains basalts of the Newer Volcanic Province, Victoria, Australia. Geochimica et Cosmochimica Acta 61 (1), 171-192.

Rencher, A. C. (2002). Methods of Multivariate Analysis. Second edition, Wiley-Interscience, New York, 708 p.

Rolland, Y., C. Picard, A. Pecher, H. Lapierre, D. Bosch, and F. Keller (2002). The Cretaceous Ladakh arc of NW Himalaya-slab melting and melt-mantle interaction during fast northward drift of Indian Plate. Chemical Geology 182 (2/4), 139-178.

Velasco-Tapia F., M. Guevara M, and S. P. Verma (2001). Evaluation of concentration data in geochemical reference materials. Chem Der Erde 61 (1), 69-91.

Verma, S. P. (1997). Sixteen statistical tests for outlier detection and rejection in evaluation of International Geochemical Reference Materials: example of microgabbro PM-S. Geostandards Newsletter: Journal of Geostandards and Geoanalysis 21 (1), 59-75.

Verma, S. P. (2005). Estadística Básica para el Manejo de Datos Experimentales: Aplicación en la Geoquímica (Geoquimiometría). Universidad Nacional Autónoma de México: México, D.F., 186 p.

Verma, S. P. (2012). Geochemometrics. Revista Mexicana de Ciencias Geológicas 29 (1), 276-298.

Verma, S.P. and L. Díaz-González (2012). Application of the discordant outlier detection and separation system in the geosciences. International Geology Review 54 (3), 593-614.

Verma, S. P. and A. Quiroz-Ruiz (2006). Critical values for six Dixon tests for outliers in normal samples up to sizes 100, and applications in science and engineering. Revista Mexicana de Ciencias Geológicas 23 (2), 133-161.

Verma, S. P., A. Quiroz-Ruiz, and L. Díaz-González (2008). Critical values for 33 discordancy test variants for outliers in normal samples up to sizes 1000, and applications in quality control in Earth Sciences. Revista Mexicana de Ciencias Geológicas 25 (1), 82-96.

Verma, S. P., L. Díaz-González, and R. González-Ramírez (2009). Relative efficiency of single-outlier discordancy tests for processing geochemical data on reference materials and application to instrumental calibration by a weighted least-squares linear regression model. Geostandards and Geoanalytical Research 33 (1), 29-49.

Verma, S. P., K. Pandarinath, S. K. Verma, and S. Agrawal (2013). Fifteen new discriminant-function-based multi-dimensional robust diagrams for acid rocks and their application to Precambrian rocks. Lithos 168-169, 113-123.

Wilks, S. S. (1963). Multivariate statistical outliers. Sankhyã A15, 407-426.

Williams, D. C. (2011). Finite sample correction factors for several simple robust estimators of normal standard deviation. Journal of Statistical Computation and Simulation 81 (11), 1697-1702.