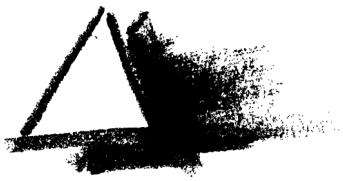


CoDa Association

CoDaWork2019

Proceedings

CoDaWork 2019

Terrassa, 3-8 June 2019

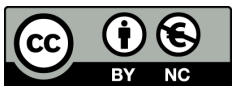
To cite this whole document, according the citation style ISO 690:1987 (UNE 50-104-94), use:

Proceedings of the 8th International Workshop on Compositional Data Analysis (CoDaWork2019): Terrassa, 3-8 June, 2019. J.J. Egozcue, J. Graffelman and M.I. Ortego (Editors). Universitat Politècnica de Catalunya-BarcelonaTECH, 2019. 202 p. ISBN 978-84-947240-2-2.

To cite a specific contribution, according the same style, use, for example:

J. Morais and C. Thomas-Agnan. Covariates impacts in compositional models and simplicial derivatives. Proceedings of the 8th International Workshop on Compositional Data Analysis (CoDaWork2019): Terrassa, 3-8 June, 2019. J.J. Egozcue, J. Graffelman and M.I. Ortego (Editors). Universitat Politècnica de Catalunya-BarcelonaTECH, 2019, p. 4–10. ISBN 978-84-947240-2-2.

ISBN: 978-84-947240-2-2



Proceedings of the 8th International Workshop on Compositional Data Analysis

Juan José Egozcue, Jan Graffelman and M.I. Ortego (Editors)

Sponsored by:

Technical support:

Welcome to CoDawork 2019

CoDaWork 2019, is the 8th international Workshop on Compositional Data analysis, and offers a forum of discussion for people concerned with the statistical treatment and modelling of compositional data or other constrained data sets, and the interpretation of models or applications involving them. The primary goal of the workshop is to identify important potential lines of future research and gain insight as to how they might be tackled.

CoDaWork 2019 intends to bring together specialist researchers, data analysts, master students, PhD students, academic scholars, as well as those with a general interest in the field, to summarize and share their contributions and recent developments.

This edition of the workshop has been organized jointly by the COSDA-UPC research group and the CoDa Association. We acknowledge work of all organizers and the support of our hosts and sponsors and promotors, in particular the city council of Terrassa, the International Association for Mathematical Geosciences (IAMG), the Universitat Politècnica de Catalunya, the Statistical Modelling Society (SMS), the Societat Catalana d'Estadística (SoCE) and the journal SORT published by the Statistical Institute of Catalonia. We also want to acknowledge the technical support from CaminsTech.

We wish you a pleasant and inspiring workshop in Terrassa!

Terrassa, June 1, 2019

Juanjo Egozcue, Chair Scientific Committee
Jan Graffelman, Chair Scientific Committee
Maribel Ortego, CoDaWork2019 Chair

CoDaWork2019 Committees

Maribel Ortego, CoDaWork2019 Chair (Universitat Politècnica de Catalunya-BarcelonaTECH)

Scientific Committee

Chairs:
Jan Graffelman and Juan José Egozcue (Universitat Politècnica de Catalunya-BarcelonaTECH)

Members:

- Cajo ter Braak, Dr., Wageningen University, Wageningen, The Netherlands
- Antonella Buccianti, Dr., Università degli Studi, Firenze, Italy
- Karel Hron, Palacky University of Olomouc, Czech Republic;
- Josep Antoni Martín-Fernández, Dr., Universitat de Girona, Girona, Spain
- Glòria Mateu-Figueras, University of Girona, Spain;
- Javier Palarea-Albaladejo, Biomathematics & Statistics Scotland, UK;
- Vera Pawlowsky-Glahn, University of Girona, Spain;
- Raimon Tolosana-Delgado, Helmholtz-Zentrum, Dresden-Rossendorf, Institute Freiberg for Resource Technology, Dresden, Germany

Local Committee:

Chair: M.I. Ortego, Universitat Politècnica de Catalunya-BarcelonaTECH, Spain.
Members:

- Jesús Corral-López, Universitat Politècnica de Catalunya-BarcelonaTECH, Spain;
- Manuel García León, Universitat Politècnica de Catalunya-BarcelonaTECH
- Manel Grifoll, Universitat Politècnica de Catalunya-BarcelonaTECH, Spain;
- Eusebi Jarauta-Bragulat, Universitat Politècnica de Catalunya-BarcelonaTECH, Spain;
- Jue Lin-Ye, Universitat Politècnica de Catalunya-BarcelonaTECH
- Glòria Mateu-Figueras, University of Girona, Spain;
- Agustí Pérez-Foguet, Universitat Politècnica de Catalunya-BarcelonaTECH, Spain;

Contents

17. Pawlowsky-Glahn, V., Planes-Pedra, M. and Egozcue, J.J. Independence test for compositional tables.

18. Quispe-Coica, F.A. and Pérez-Foguet, A. Joint evolution of access to water of urban and rural populations in South America through Compositional Data Analysis.

19. Sánchez-Balseca, J. and Pérez-Foguet, A. Assessing CoDa regression for modelling daily multivariate air pollutants evolution.

20. Srakar, A. and Fry, T.R.I. Wavelet regressions for compositional data.

21. Todorov ,V. Monitoring robust estimates for compositional data.

22. Tolosana-Delgado, R., Talebi, H., Khodadadzadeh, M. and van den Boogaart, K.G. On machine learning algorithms and compositional data.

23. Vermeesch, P. Statistical models for point-counting data.

24. Ziembik, Z. and Dolhanczuk-Sródka, A. Application of multivariate imputation of left-censored data in biomonitoring of radioisotopes.

Covariates impacts in compositional models and simplicial derivatives

J. Morais[1,2], and C. Thomas-Agnan[2]
[1]Avisia, Bordeaux, France; *joanna.morais@live.fr*
[2]Toulouse School of Economics, Toulouse, France

Summary

Compositions can be used as variables in regression models, either as explanatory variables (see Hron et al. (2012)) or as dependent variables (see Egozcue et al. (2012)), or both (see Chen et al. (2016), Morais et al. (2018b) and Nguyen T.H.A (2018)). However, measuring the marginal impacts of covariates in these types of models is not straightforward, as the change in one component of a composition may affect the rest of the composition.

Morais et al. (2018a) have shown how to measure, compute and interpret these marginal impacts in the case of linear regression models with a dependent composition (Y) by compositional explanatory variables (X). The resulting natural interpretation is in terms of an elasticity, commonly used in econometrics and marketing applications. Morais et al. (2018a) also demonstrate the link between these elasticities and simplicial derivatives as defined in Egozcue et al. in Pawlowsky-Glahn and Buccianti (2011), chapter 12 and Barcelo-Vidal et al. in Pawlowsky-Glahn and Buccianti (2011), chapter 13.

The aim of this contribution is to show how to compute these semi-elasticites and simplicial derivatives in other situations, namely first when the dependent variable is a composition and the explanatory variables are non-compositional, and second when the dependent variable is non-compositional and at least one of the explanatory variables is a composition. Moreover we also consider the case where a total is used or not as an explanatory variable, with several possible interpretations of the total.

Finally, we discuss how to compute confidence intervals for these elasticities or semi-elasticities, which significantly improves the interpretability of the compositional regression models. This contribution will be illustrated by real-data applications.

Key words: compositional regression model, interpretation, simplicial derivative, elasticity

1 Introduction

Compositional regression models have been investigated from a theoretical perspective, for example in the following books: Pawlowsky-Glahn and Buccianti (2011), Van Den Boogaart and Tolosana-Delgado (2013), and Pawlowsky-Glahn et al. (2015). However, few articles are applying them in practice. Hron et al. (2012) present a case where the explanatory variables are compositional (called 'X-compositional' model below). Egozcue et al. (2012) focus on the case where the dependent variable is a composition (called 'Y-compositional' model below).

The case where a compositional dependent variable is explained by component-dependent explanatory variables (called 'YX-compositional model' below) has been addressed in quite recent articles: Wang et al. (2013), Kynclova et al. (2015), Chen et al. (2016), Morais et al. (2018b), Morais et al. (2018a) and Morais

et al. (2017). However, Wang et al. (2013) presents a simplified model compared to the others, which has not been mentioned in the books we cite above. As shown in Morais et al. (2018b), this model is equivalent to what is called MCI (multiplicative competitive interaction) model in the econometric / marketing literature.

2 Covariate impact in compositional models

2.1 Compositional model specifications

We are going to consider the impacts of covariates in the following three main types of compositional regression models (see Table 1):

- Y-compositional model: the dependent variable is a composition and the explanatory variables are non-compositional
- X-compositional model: the dependent variable is non-compositional and at least one of the explanatory variables is a composition
- YX-compositional model: the dependent variable is a composition and at least one of the explanatory variables is a composition

Table 1: Compositional models and notations

	Y-compositional model	X-compositional model	YX-compositional model
Specification in $\mathcal{S}^D$	$\mathbf{Y}_t = \mathbf{a} \oplus \check{X}_t \odot \mathbf{b} \oplus \epsilon_t$ $\oplus T(\check{\mathbf{Y}})_t \odot \mathbf{c}$	$\check{Y}_t = a + < \mathbf{b}, \mathbf{X}_t >_A + \epsilon_t$ $+ cT(\check{\mathbf{X}})_t$	Model 'CODA': $\mathbf{Y}_t = \mathbf{a} \oplus \mathbf{B} \boxdot \mathbf{X}_t \oplus \epsilon_t$ $\oplus T(\check{\mathbf{X}})_t \odot \mathbf{c}$ Model 'MCI': $\mathbf{Y}_t = \mathbf{a} \oplus b \odot \mathbf{X}_t \oplus \epsilon_t$ $\oplus T(\check{\mathbf{X}})_t \odot \mathbf{c}$
Specification in $\mathbb{R}^{D-1}$	$\mathbf{Y}_t^* = \mathbf{a}^* + \mathbf{b}^* \check{X}_t + \epsilon_t^*$ $+ \mathbf{c}^* T(\check{\mathbf{Y}})_t$	$\check{Y}_t = a + \sum_{k=1}^{D_X - 1} b_k^* X_{t,k}^* + \epsilon_t$ $+ cT(\check{\mathbf{X}})_t$	Model 'CODA': $\mathbf{Y}_t^* = \mathbf{a}^* + \mathbf{B}^* \mathbf{X}_t^* + \epsilon_t^*$ $+ \mathbf{c}^* T(\check{\mathbf{X}})_t$ Model 'MCI': $\mathbf{Y}_t^* = \mathbf{a}^* + b\mathbf{X}_t^* + \epsilon_t^*$ $+ \mathbf{c}^* T(\check{\mathbf{X}})_t$
Notations	$\mathbf{Y}_t, \mathbf{a}, \mathbf{b}, \epsilon_t \in \mathcal{S}^{D_Y}$, $\check{X}_t \in \mathbb{R}$ $\mathbf{Y}_t^*, \mathbf{a}^*, \mathbf{b}^*, \epsilon_t^* \in \mathbb{R}^{D_Y - 1}$	$\mathbf{X}_t, \mathbf{b} \in \mathcal{S}^{D_X}$, $\check{X}_t, a, \epsilon_t \in \mathbb{R}$ $\mathbf{X}_t^*, \mathbf{b}^* \in \mathbb{R}^{D_X - 1}$	$\mathbf{B} \in \mathbb{R}^{D_Y, D_X}$, $b \in \mathbb{R}$ $\mathbf{B}^* \in \mathbb{R}^{D_Y - 1, D_X - 1}$

 ISBN 978-84-947240-2-2

Let us denote by $\check{\mathbf{X}}_\mathbf{t} = (\check{X}_{1t}, \cdots, \check{X}_{D_X t})' \in \mathbb{R}_+^{D_X}$ a vector of so-called volumes for a variable X, with $\sum_{i=1}^{D_X} \check{X}_{it} = T(\check{\mathbf{X}}_\mathbf{t})$, and $\mathbf{X}_t = \mathcal{C}(\check{X}_{1t}, \cdots, \check{X}_{D_X t})' = \left(\frac{\check{X}_{1t}}{T(\check{\mathbf{X}}_\mathbf{t})}, \cdots, \frac{\check{X}_{D_X t}}{T(\check{\mathbf{X}}_\mathbf{t})}\right)' = (X_{1t}, \cdots, X_{D_X t})' \in \mathcal{S}^{D_X}$ the corresponding composition carrying the relative information, with $\sum_{i=1}^{D_X} X_{it} = 1$.

In subsections 2.2 and 2.3 we define and give formulas and interpretations for elasticities and semi-elasticities in these models. Then in section 3, we explain how these quantities are linked to simplicial derivatives, which is the reason why they are the natural tool for interpreting impacts in these models.

2.2 Elasticities for YX-compositional models

Morais et al. (2018a) have shown how to interpret the relative impact of an X component on a Y component (without total). The resulting interpretation uses an elasticity, commonly found in econometric applications:

$$E(\mathbf{Y}_t, \check{\mathbf{X}}_\mathbf{t})_{D_Y \times D_X} = \frac{\partial \log \mathbb{E}^{\oplus} \mathbf{Y}_t}{\partial \log \check{\mathbf{X}}_\mathbf{t}}$$

The interpretation is as follows: the relative impact of an increase of 1% of the volume of the j^{th} component $\check{X}_{jt}$ on the component Y_{it} is equal to E_{ij}%. Note that since $\mathbb{E}^{\oplus}\mathbf{Y}_t$ is a scale-invariant function of $\mathbf{X}_t$, the result is also scale-invariant. To compute the impact on the volumes $\check{Y}$, one may multiply the above elasticity by $T(\check{\mathbf{Y}}_\mathbf{t}) = \sum_{i=1}^{D_Y} \check{Y}_{it}$.

2.3 Semi-elasticities for Y-compositional models and X-compositional models

In the case of Y-compositional and X-compositional models, the natural tool is semi-elasticities. However the formulas differ in the two cases:

- Y-compositional case: $SE(\mathbf{Y}_t, \check{X}_t) = \frac{\partial \log \mathbb{E}^{\oplus} \mathbf{Y}_\mathbf{t}}{\partial \check{X}_t}$
- X-compositional case: $SE(\check{Y}_t, \check{\mathbf{X}}_t) = \frac{\partial \check{Y}_t}{\partial \log \check{\mathbf{X}}_t}$

In the Y-compositional case, the interpretation is as follows: the relative impact of an additive increase of 1 unit of the volume of $\check{X}_t$ on the component Y_{it} is equal to SE_i%.

In the X-compositional case, the interpretation is as follows: the additive impact of an increase of 1% of the volume of the component $\check{X}_{jt}$ on the dependent variable $\check{Y}_t$ is equal to SE_j units.

2.4 Composition total as explanatory variable

In some cases, it may be relevant to also include in the model a variable measuring a total (not scale-invariant). If the composition $(X_1, \cdots, X_D)$ is the closure of a vector of non-constant sum volumes (positive numbers) $(\check{X}_1, \cdots, \check{X}_D)'$, then Pawlowsky-Glahn et al. (2015) argue that different formulas can be used for this total:

- Arithmetic total: $T(\check{\mathbf{X}}) = \sum_{i=1}^{D} \check{X}_i$
- Geometric total: $T(\check{\mathbf{X}}) = (\prod_{i=1}^{D} \check{X}_i)^{1/\sqrt{D}}$

The presence of this total variable has to be taken into account in the partial impact measure computations.

2.5 Computation of elasticities and semi-elasticities

Table 2 presents the semi-elasticities and elasticities expressions for the three types of models, with or without a compositional total. In the presence of a total, as we will explain in Section 3, we need to distinguish three types of impacts:

- Type 1 impact: when the total remains constant and we look at derivatives in the direction of one of the unitary vectors of an orthonormal basis of $\mathcal{S}^{D_X}$. With a proper choice of basis and of contrast matrix as in Hron et al. (2012), this corresponds to an infinitesimal change in one component keeping all but the first ILR constant.
- Type 2 impact: when the composition remains constant and we look at ordinary derivatives with respect to the total
- Type 3 impact: when one of the components varies together with the total.

Table 2: (Semi-)elasticities, without and with the total

		Y-compositional model	X-compositional model	YX-compositional model
WITHOUT TOTAL		$SE(\mathbf{Y_t}, \check{X}_t) = \frac{\partial \log \mathbb{E}^{\oplus} \mathbf{Y_t}}{\partial \check{X}_t}$ $= \mathbf{W}_t^* \mathbf{b}^*$ $= \mathbf{W}_t \log \mathbf{b}$	$SE(\check{Y}_t, \check{\mathbf{X}}_t) = \frac{\partial \mathbb{E}(\check{Y}_t)}{\partial \log \check{\mathbf{X}}_t}$ $= \mathbf{V}\mathbf{b}^*$ $= \mathbf{V}\mathbf{V}' \log \mathbf{b}$	Model 'CODA': $E(\mathbf{Y_t}, \check{\mathbf{X}}_t) = \frac{\partial \log \mathbb{E}^{\oplus} \mathbf{Y_t}}{\partial \log \check{\mathbf{X}}_t}$ $= \mathbf{W}_t^* \mathbf{B}^* \mathbf{V}' = \mathbf{W}_t \mathbf{B}$ Model 'MCI': $E(\mathbf{Y_t}, \check{\mathbf{X}}_t) = \mathbf{W}_t b$
WITH TOTAL	**Type 1**	Like without total	Like without total	Like without total
	Type 2	No meaning	$ME(\check{Y}_t, T(\check{\mathbf{X}})_t)$ $= \frac{\partial \check{Y}_t}{\partial T(\check{\mathbf{X}})_t} = c$	Like Y-compositional $SE(\mathbf{Y}_t, T(\check{\mathbf{X}})_t) = \mathbf{W}_t^* \mathbf{c}^*$
	Type 3	$SE(\mathbf{Y_t}, T(\check{\mathbf{Y}})_t)$ $= \mathbf{W}_t^* \mathbf{c}^*$ $= \mathbf{W}_t \log \mathbf{c}$	$SE(\check{Y}_t, \check{\mathbf{X}}_t)$ $= \mathbf{V}\mathbf{b}^* + c \frac{\partial T(\check{\mathbf{X}})_t}{\partial \log \check{\mathbf{X}}_t}$	$E(\mathbf{Y_t}, \check{\mathbf{X}}_t) =$ $\mathbf{W}_t^* \left(\mathbf{B}^* \mathbf{V}' + \log \mathbf{c} \frac{\partial T(\check{\mathbf{X}})_t}{\partial \log \check{\mathbf{X}}_t} \right)$ $= \mathbf{W}_t \left(\mathbf{B} + \log \mathbf{c} \frac{\partial T(\check{\mathbf{X}})_t}{\partial \log \check{\mathbf{X}}_t} \right)$

Notations $\mathbf{W}_{t_{(D_Y, D_Y)}}$ is a matrix with $(1 - Y_{it})$ on the diagonal and $-Y_{it}$ elsewhere on the ith row.

These results are based on the following lemmas. Lemma 2.1 computes a semi-log derivative of an ILR transformation and Lemma 2.2 a semi-log derivative of the inverse of an ILR transformation.

Lemma 2.1 *If* $\mathbf{z}$ *is a D-composition which is the closure of the vector* $\check{\mathbf{z}}$ *of* $\mathbb{R}^D_+$*, and if* $\mathbf{z}^* = ilr(\mathbf{z}) = \mathbf{V}'log(\mathbf{z})$ *is the ILR-transformed composition associated to the contrast matrix* $\mathbf{V}$*, then*

$$\frac{\partial ilr(\mathbf{z})}{\partial \log \check{\mathbf{z}}} = \mathbf{V}'$$

Lemma 2.2 *If* $\mathbf{z}$ *is a D-composition which is the closure of the vector* $\check{\mathbf{z}}$ *of* $\mathbb{R}^D_+$*, and if* $\mathbf{z}^* = ilr(\mathbf{z}) = \mathbf{V}'log(\mathbf{z})$ *is the ILR-transformed composition associated to the contrast matrix* $\mathbf{V}$*, then*

$$\frac{\partial \log(\check{\mathbf{z}})}{\partial \log \mathbf{z}^*} = \mathbf{W}\mathbf{V},$$

where $\mathbf{W}$ *is the* $D \times D$ *matrix with* $(1 - z_{it})$ *on the diagonal and* $-z_{it}$ *elsewhere on the ith row.*

The results of Table 2 are obtained combining, according to each case, Lemma 2.1 (for $\mathbf{z} = \mathbf{X}$) and Lemma 2.2 (for $\mathbf{z} = \mathbf{Y}$) with the marginal effects of the model specified in $\mathbb{R}^{D-1}$.

3 Semi-elasticities and simplicial partial derivatives

Morais et al. (2018a) have shown how to compute elasticities for YX-compositional model, and how they are linked to simplicial derivatives.

In the case of Y-compositional and X-compositional models, we can compute semi-elasticities of a composition relative to a non-compositional variable, and of a non-compositional variable relative to a composition, and we can show that they are linked to the simplicial derivatives using (Egozcue et al. in Pawlowsky-Glahn and Buccianti (2011), chapter 12, and Barcelo-Vidal et al. in Pawlowsky-Glahn and Buccianti (2011), chapter 13), as presented in Table 3.

Indeed, in the case of the X-compositional model, Propositions (13.10) and (13.13) in Barcelo-Vidal et al. in Pawlowsky-Glahn and Buccianti (2011), chapter 13, imply that the part-$\mathcal{C}$ derivatives of an homogeneous function of degree zero $\underline{f}$ of a composition $\mathbf{x} = \mathcal{C}(\check{\mathbf{x}})$, which we denote here by $\frac{\partial \underline{f}(\mathbf{x})}{\partial^{\oplus}\mathbf{x}}$ is given by:

$$\frac{\partial \underline{f}(\mathbf{x})}{\partial^{\oplus}\mathbf{x}} = \frac{\partial f(\check{\mathbf{x}})}{\partial \log(\check{\mathbf{x}})}$$

Therefore the derivative of a function $\underline{f}$ of a simplex valued variable $\mathbf{x} = \mathcal{C}(\check{\mathbf{x}})$ corresponds to the ordinary semi-log derivative of the corresponding homogeneous function f of the volumes $\check{\mathbf{x}}$.

Similarly, in the case of the Y-compositional model, for a simplex-valued function f of a real variable $\check{x} \in \mathbb{R}^+$, Theorem 12.2.6 in Egozcue et al. in Pawlowsky-Glahn and Buccianti (2011), chapter 12, implies that:

$$\frac{\partial^{\oplus} f(\check{x})}{\partial \check{x}} = \mathcal{C}\left(\exp\left(\frac{\partial \log f(\check{x})}{\partial \check{x}}\right)\right)'$$

This result links the derivatives of a simplex-valued function f to the semi-log derivatives (in the ordinary sense) of the function f.

Finally, considering models including a total, one would need to define infinitesimal paths in the $\mathcal{T}$-space. Instead we consider three types of infinitesimal variations as described in Section 2.5. For Type 1 and 2, no new theory is needed: Type 1 corresponds to derivatives with respect to a simplex valued variable and Type 2 to ordinary derivatives with respect to the total. For Type 3, the easiest is to express the dependent as a function of the volumes and use ordinary derivatives of the ensuing function.

Table 3: Simplicial derivative and (semi-)elasticities

	Y-compositional model	X-compositional model	YX-compositional model
Simplicial derivative in $\mathcal{S}^D$	$\frac{\partial^{\oplus}\mathbb{E}^{\oplus}\mathbf{Y}_t}{\partial X_t} = \mathcal{C}\left(\exp\left(\frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y}_t}{\partial X_t}\right)\right)' = \mathcal{C}\left(\exp SE(\mathbf{Y}_t, X_t)\right)'$	$\frac{\partial \mathbb{E}Y_t}{\partial^{\oplus}\mathbf{X}_t} = \left[\frac{\partial \mathbb{E}Y_t}{\partial \log \check{X}_{jt}}\right] = SE(\check{Y}_t, \check{\mathbf{X}}_t)$	$\frac{\partial^{\oplus}\mathbb{E}^{\oplus}\mathbf{Y}_t}{\partial^{\oplus} X_{jt}} = \mathcal{C}\left(\exp\left(\frac{\partial \log \mathbb{E}^{\oplus}\mathbf{Y}_t}{\partial \log \check{X}_{jt}}\right)\right)' = \mathcal{C}\left(\exp E(\mathbf{Y}_t, \check{\mathbf{X}}_t)\right)'$

4 Illustration

This application aims to explain the Body Mass Index (BMI) of Chinese individuals by the following characteristics: their age, gender, ethnicity, education, physical activity, size of their household, along with their food consumption composition in terms of macronutrients: calories in carbohydrate ($Kcal_C$), fat ($Kcal_F$) and protein ($Kcal_P$). Thus, it is a X-compositional model. This model can be specified as follows:

$$\begin{aligned} Y_t = a &+ b_1^* ilr1_t + b_2^* ilr2_t \\ &+ c_1 \log(Age_t) + c_2 \log(Age_t)^2 + c_3 Gender_t + c_4 Ethnic_t \\ &+ c_5 EducUniv_t + c_6 PhysHeavy_t + c_7 Hsize_t + \epsilon_t \end{aligned}$$

with $ilr1, ilr2$ the ILR coordinates of the composition $\mathbf{Kcal} = \mathcal{C}(Kcal_C, Kcal_P, Kcal_F)'$ and $Y = \log(BMI)$. The parameters estimated by least squares are:

$$(b_1^*, b_2^*)' = (-0.0176, 0.0108)' \Longleftrightarrow (b_C, b_P, b_F)' = (0.3285, 0.3383, 0.3332)'$$

Then, the vector of semi-elasticities of Y relative to an infinitesimal relative variation of the components of **Kcal** is $(-0.0144, 0.0149, -0.0005)'$, meaning that an increase of 1% of the carbohydrate kilo-calories, keeping the consumption of fat and protein unchanged, results in an approximate 0.014 units decrease of the $\log(BMI)$. If, in the same model, the arithmetic total is included, then this semi-elasticity is even stronger: 0.016 instead of 0.014, but with a geometric total, the semi-elasticity is lower: -0.010.

5 Conclusion

This contribution explains how to interpret all types of compositional regression models using the well-adapted (semi-)elasticities. Confidence intervals on (semi-)elasticities can be computed by the Delta method, or simply using a bootstrap approach. Further work have to be done on this part.

Acknowledgements

We thank Trinh Thi Huong for giving us access to the data and her R codes that we used for the illustration.

References

Chen, J., X. Zhang, and S. Li (2016). Multiple linear regression with compositional response and covariates. *Journal of Applied Statistics*, 1–16.

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*.

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

Kynclova, P., P. Filzmoser, and K. Hron (2015). Modeling compositional time series with vector autoregressive models. *Journal of Forecasting 34*(4), 303–314.

Morais, J., C. Thomas-Agnan, and M. Simioni (2017). Impact of advertising on brand's market-shares in the automobile market: a multi-channel attraction model with competition and carryover effects.

Morais, J., C. Thomas-Agnan, and M. Simioni (2018a). Interpretation of explanatory variables impacts in compositional regression models. *Austrian Journal of Statistics 47*(5), 1–25.

Morais, J., C. Thomas-Agnan, and M. Simioni (2018b). Using compositional and dirichlet models for market share regression. *Journal of Applied Statistics 45*(9), 1670–1689.

Nguyen T.H.A, Laurent T., T.-A. C. R.-G. A. (2018). Analyzing the impacts of socio-economic factors on french departmental elections with coda methods. *TSE Working Paper 18*(961).

Pawlowsky-Glahn, V. and A. Buccianti (2011). *Compositional data analysis: Theory and applications*. John Wiley & Sons.

Pawlowsky-Glahn, V., J. J. Egozcue, and D. Lovell (2015). Tools for compositional data with a total. *Statistical Modelling 15*(2), 175–190.

Pawlowsky-Glahn, V., J. J. Egozcue, and R. Tolosana-Delgado (2015). *Modeling and Analysis of Compositional Data*. John Wiley & Sons.

Van Den Boogaart, K. G. and R. Tolosana-Delgado (2013). *Analysing Compositional Data with R*. Springer.

Wang, H., L. Shangguan, J. Wu, and R. Guan (2013). Multiple linear regression modeling for compositional data. *Neurocomputing 122*, 490–500.

ISBN 978-84-947240-2-2

Is Municipal Solid Waste composition affected by Demographics or Seasons?

John Chandler
A.J. Chandler & Associates Ltd. *john.chandler@bell.net*

Summary

Residual wastes, material remaining after recycling and organics diversion, require disposal. One alternative is to create Solid Recovered Fuels [SRF] for energy intensive industries. Frequently, potential users of SRF require that the material be fully characterised come with guarantees of energy values and contamination levels. With limited data on compositional variations of residual wastes, particularly as it may pertain to an SRF product, a municipality in southern Ontario, Canada, required a detailed study of the composition and quality of their residual wastes. The wastes are collected separately from both single-family [CS] and high-rise residential [MR] properties and the study afforded an opportunity to distinguish between these waste streams, as well as addressing seasonal differences by sorting materials at different times of the year.

Changing priorities during the study resulted in differences in the sorting categories for the 2 phases reported here. Moreover, components that were common to both phases showed different quantities by source and season. The question arose, were the differences significant? Not unexpectedly, the sort data was not normally distributed and applying classical statistical techniques was not successful. Enter, compositional data analysis techniques. When applying these techniques, it became evident that working with high dimensional data increases the complexity of the analyses and the difficulty of interpreting the results. This extended abstract covers some of the methods used to date. From these it was concluded that the dimensionality of the data should be further reduced to attempt to address the question raised above. The final presentation will include the analysis of the less complex data set.

1 Introduction

A team from the University of Waterloo helped the municipality develop a detailed waste composition study protocol, Chandler et al. (2017), aimed at gathering data on a wide range of materials in the waste streams. The sorting list reflected materials that can be diverted in the provincial blue box program (45), and the provincial MHSW initiative (9) as well as non-recyclable materials. During Phase 1, the waste was screened to 4 size fractions: +50 mm; 25-50 mm; 6-25 mm; minus 6 mm. The 25-50 mm fraction being classified into 8 broad categories. For Phase 2, the +50 mm fraction was split into +100 mm and 50-100 mm size fractions, but the overall number of categories was reduced to 32 with 29 being identified in both the larger size fractions. Moisture, heating value, and trace metal and chlorine levels were defined as important characteristics for laboratory investigation. The fines fractions were tested for their suitability as feed to anaerobic digestion processes. In each phase the objective was to separate sixteen 100 kg samples from the bulk of material arriving from each source. In Phase 1, fall, only 28 samples were completed and sorted into 43 parts; in Phase 2,

winter, 31 samples were sorted into 62 parts. An additional sample was sort but into a reduced number of parts, so it was not considered for the initial analyses.

2 Methods

With four distinct data sets, the high-dimensional structure limited the ability to define outliers and perform regression analysis with robust techniques. To compare the source differences by phase another approach was necessary. Moreover, if the seasonal differences were to be examined, the parts had to be aligned for both phases.

Some parts contained zero quantities for some sorts. The missing data were classified as Rounded Zeros, assuming they would have been found in larger samples. Before applying imputation methods though, it was necessary to recognize that, for quality imputation of replacement values, 50% zeros should the maximum used, Martin-Fernandez (2019). To address this, two parts with more than 80% zeros were removed from the compositions. The remaining part with greater than 50% zeros were combined with similar materials to create 32 components for Phase 1. The result was between 6.0% and 7.5% zeros in the data. Phase 2 required similar aggregating. Some of the 50-100 mm parts were combined with their larger counterparts to reduce zeros. In addition, parts with similar characteristics were combined to reduce the list to 43 parts. The 31 tests contained 4.2 to 6.4% zeros before imputation.

The list of parts for Phase 1 and Phase 2 was not identical, and direct comparisons to identify seasonal effects in waste from the two sources required more adjustments. The raw data was processed as second time. The two larger size fractions in Phase 2 were combined to replicate the +50 mm size range of Phase 1. Plastics, metals, textiles and other parts were combined so both phases had the same categories. Furthermore, three parts: Yard; MHSW; and Ceramics, were dropped from consideration due to high zero counts. The resulting 19 parts could be compared for seasonal effects. Reducing the list of parts to 19 resulted in limited missing data: 6-7% for Phase 1 and 2.5-3% for Phase 2.

Consolidation of the data sets was completed using the original data, and then each data set had the zeros replaced using an imputation technique. With more parts than observations, the lrEM procedure did not work and Martin-Fernandez suggested that the multKM algorithm be applied. The threshold used was the smallest measured value for each part of each source regardless of D. Imputation for the phase data was also conducted using the minimum value from either source. After imputation, the data sets were closed.

Cluster analysis provides a visual illustration of potential similarities and differences between data, and is not limited by the relationship between D and n. Cluster analyses used a hierarchical agglomeration procedure, *hclust,* on the clr-transformed data. A second clustering, Q-mode, utilizes the variation matrix converted to a distance matrix was used to detect patterns of variation between parts. For the initial pass, the high dimension data were used to create dendrograms for each source and each season as well as for each phase to allow the combined source data to be examined. A similar exercise was completed for the reduced dimensional data, D=19.

Since the dendrograms suggested that there could be outliers in the data, an outlier identification process was initiated. Unfortunately, the outCoDa function in *robCompositions* did not work for the initial high-dimensional data sets for the sources,

even when the samples from each phase were examined together. When performed on the D=19 data sets, it was still necessary to examine larger data sets: all CS and the all MR samples regardless of season; and, the combined seasonal sort data. The results still indicated caution would be necessary in interpretation due to limited sample numbers. Another reduction in the number of parts will be completed to enable more definition of outliers.

Filzmoser et al. (2019) discuss the difficulties that arise when trying to analyse high-dimensional compositions and recommend Partial Least Squares [PLS] for regression and classification of such data. Since the objective of this study was to determine if there were differences between sources and seasons, the two-group classification system suggested by these authors was applied. To find a model that can classify the parts into either of the source groups and identify which parts are "significantly" different in the two groups, the PLS procedure computes values using the clr coefficients and that the variance of the regression coefficients is estimated by a jackknife procedure. The procedure allows the "optimal" number of PLS components to determined. From that point, it is possible to infer the regression coefficients for the model based upon single clr coefficients and determine which parts are significant in the model.

Pairwise logratios can be used to identify parts to distinguish between the 2 source groups. This method, Walach et al. (2017), employs the variation matrix elements of all observations jointly and compares them to those computed for the single groups. Large differences, expressed as the statistic V_j^*, indicate potential marker variables. Filzmoser et al. note that calculating the variance in ***robCompositions*** can be completed using classical empirical variance techniques but it can also be completed using the robust τ estimator to establish the variance of the variation matrix taking into consideration outliers in the data. PLS techniques were applied to the initial high dimensional data to compare the results of source specific sorting data for each phase, and to the reduced dimensional data to compare both the source specific data and the seasonal differences between the results.

The outCoDa procedure was employed on the D=19 compositions. It is limited to considering outliers of data sets which contained at least as many samples as dimensions and will not work with the high dimensional data sets. The procedure produces a warning message along with the results unless n>2D. Filzmoser et al. caution that the procedure is more likely to identify outliers that are associated with parts that have low concentrations than dominating components. This suggests that even though the method identifies outlying samples, care should be taken in interpreting their significance.

3 Discussion

Imputation for each source and phase was based upon the lowest recorded weight for each component for the group. The approach created 6 thresholds, but it could also introduce more variability. To examine this potential cluster analyses of the phase data was used. Before discussing those differences, a quick review of the source clustering by phases. Clustering of the individual source samples in each phase showed that same day sorts were seldom combined. The sampling effort aimed to collect different samples from the material arriving at the site, and the clustering suggests that was the case. The plots contain sorts that join late in the process indicating that they might be different. When the parts are clustered Pet Waste and WEEE appear to be outliers in 3 of the 4 source related clusters. Construction waste and metal components

appeared to be outliers in Phase 1 CS and Phase 2 MR. Phase 2 resulted in Other Textiles being identified as a potential outlier for both sources. The other components that suggest outliers were: Containers with liquid; Paper fibres in large fines in Phase 1; MHSW in CS2 and
Diapers in MR2. The latter reflects some very large masses of diapers found sorts from the last day of sorting which suggests that the same care might not have been taken in sampling that day.

The dendrograms for the two phases were compared to see if there were differences that might be attributed to the threshold used. The number of groups at a given height is an indication that there may be differences. For Phase 1 at 11 there were 3 groups with the source specific threshold and 4 when the phase threshold was used; however, there were 8 sorts that were late in joining various groups in the source specific clustering and only 4 with the phase specific clustering. The pattern observed in the Phase 2 sort clustering is that there appear to be more groups with the phase specific threshold. That approach produced 6 isolated sorts while the source specific approach had only 4 isolated sorts. Generally, the groups contain the same members regardless of the threshold approach. As with the Phase 1 parts clustering, Phase 2 showed WEEE, Pet Waste, MHSW, Construction waste, synthetic textiles and yard waste as separated parts regardless of the threshold approach. Overall the parts clustering appears to be similar with only slightly different positioning that could be attributed to the threshold approach.

The objective of the PLS procedure was to classify the data into two groups, (sources), thus only the phase results were used. The analyses were run for both threshold options for the D=32 or 43 data sets, Table 1. Also included in the table is similar data for the D=19 data for the sets of combined data as labelled.

Table 1 Summary of PLS Analyses of Initial Sort Data

Phase	Group	ncomp	Significant	High Significance	V*	V*tau
1 D=32	Source	1	4, 8, 9, 12, 15, 22, 23, 24, 30	1, 3, 20, 26, 27	26, 27	26
	Phase	6	3, 9, 20	8	8, 27	8, 27
2 D=43	Source	4	2, 5, 6, 7, 27, 30, 38	17, 19, 33, 34, 39	38	32
	Phase Th.	4	2, 5, 6, 16, 27, 35	7, 17, 19, 33, 34, 38, 39	38, 32	32
1 D=19	Curbside	6		5, 16	16	16
2 D=19	Multi-Res	4	5, 6	1, 13, 16	16	16
Both D=19	Combined	2	1, 5	3, 13, 16, 17	16	16

The procedure identifies the number of components, ncomp, that provide the best classification of the data. Also identified were the parts that were significant and those that were highly significant. The marker variables were identified, V^*, and the effect on the markers of potential outliers in the data, V^*_{tau}.

For Phase 1 the phase minimum threshold data suggested 2 markers, and outliers had no influence. For the source minimum threshold, there were 2 markers but outliers reduced this to a single markers. There are differences between the markers that can be attributed to the threshold approach. For Phase 2 both alternatives resulted in 4 components that produced the best identification. The markers were similar parts but

the outliers appear to have changed them, substituting a different marker for the source threshold and reducing the markers for the other alternative.

The results of both the cluster analysis and the PLS regression suggest that the threshold and its application may influence the data. Since waste streams are likely to be different from the sources, a threshold based upon the source minimum by phase was used for the balance of the analyses.

The second step of the analysis employed the reduced compositional data set, 19 parts for both Phase 1 and Phase 2. Cluster analyses were completed for both phases. For both phase the sorts clustered in the same manner as the larger dimensional data, no specific connections by sorting day. The isolated sorts were more limited for both the CS and MR sorts; however, when combined CS and MR was clustered, Phase 1 had fewer isolated sorts (2) than Phase 2 (7). The parts clusters for CS isolated WEEE, Pet Waste, Construction, and diapers. MR clusters had WEEE, Construction, Diapers in Phase 2 and WEEE, Pet Waste, Containers with Liquid, and Personal Care products in Phase 1. Not surprisingly, the combined CS and MR for each phase isolated Construction, WEEE, and Pet Waste. These results are like those found for the larger dimensional data.

Rerun the PLS for the reduced component list provides the result shown in the bottom of Table 1. The marker component in all cases was 16, the large fines. Fewer components were identified as significant or highly significant. Again, the caution that the number of samples was less than twice the number of parts should be noted.

The outlier tests for the combined data from both phases for Curbside and Multi-Res sources. The CS data had 6 out of 30 values identified as outliers: 4, 8, 11, 17, 19, 21. These were split equally between the two tests. The MR test produced 5 outliers: 2, 6, 14, 19, 28. The first two from Phase 1 the others from Phase 2.

Biplots of the D=19 results were produced for both the CS and MR sampling, Figure 1. These include all the data collected from each source. The components that showed high variability in the cluster analyses produce long vectors in these plots. The bulk of the components have relatively short vector, but large fines are noticeable in each plot.

The final figures available currently are boxplots of the isometric log-ratio coordinates for each source and each season, Figure 2. Typically, when comparing values on boxplots, if the interquartile ranges overlap between different parameters there is little difference between them; however, should the boxes not overlap it would suggest that significant differences exist between the groups. Consider irl15 in the upper plot, none of the boxes overlap suggesting that there were differences between the sources in both seasons, and moreover even the source related boxes do not overlap suggesting that there are differences in this component between the two seasons.

Conclusions

Analyses are on-going on these data. Before the conference the author hopes to have completed the analysis using MANOVA to distinguish between the groups and quantify the significance of the observations presented in this paper.

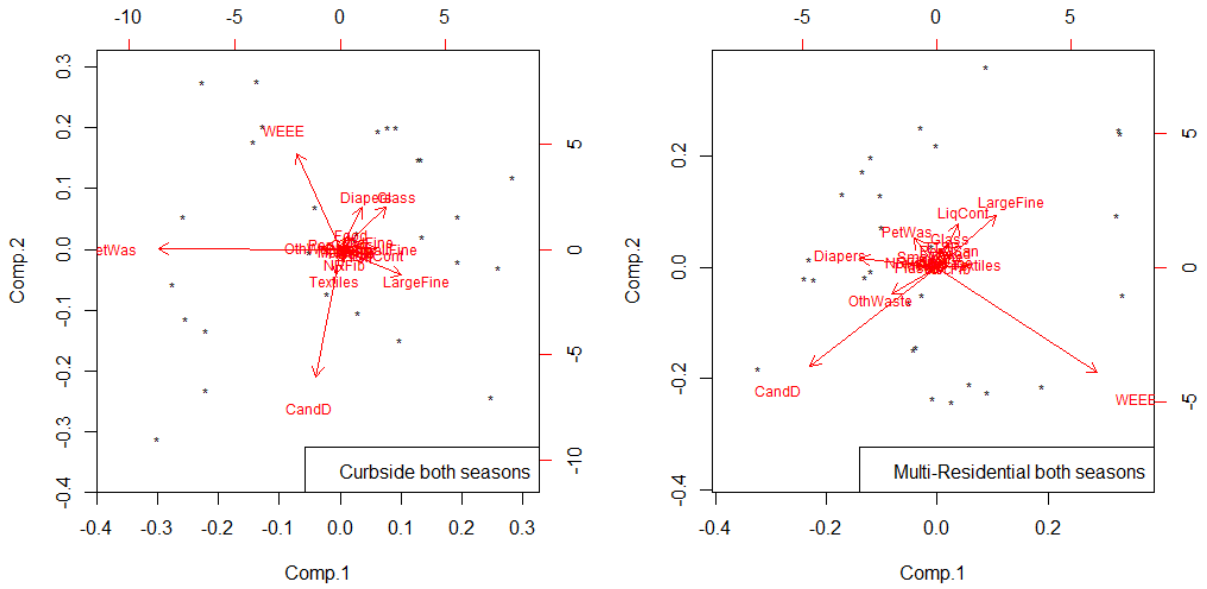
References

Martin-Fernandez, J.A., (2019) Personal Communication with author

Figure 1 Biplots for Different Sources

Filzmoser, Peter, Karel Hron, and Matthias Templ (2019). Applied Compositional Data Analysis with Worked Examples in R. Springer Series in Statistics. Chapter 11.

Walach, J., P. Filzmoser, K. Hron, and B. Walczak (2017). Robust biomarker identification based on pairwise log-ratios. Chemom. Intell. Lab. Syst. **171**, 277-285. As noted in Filzmoser et al.



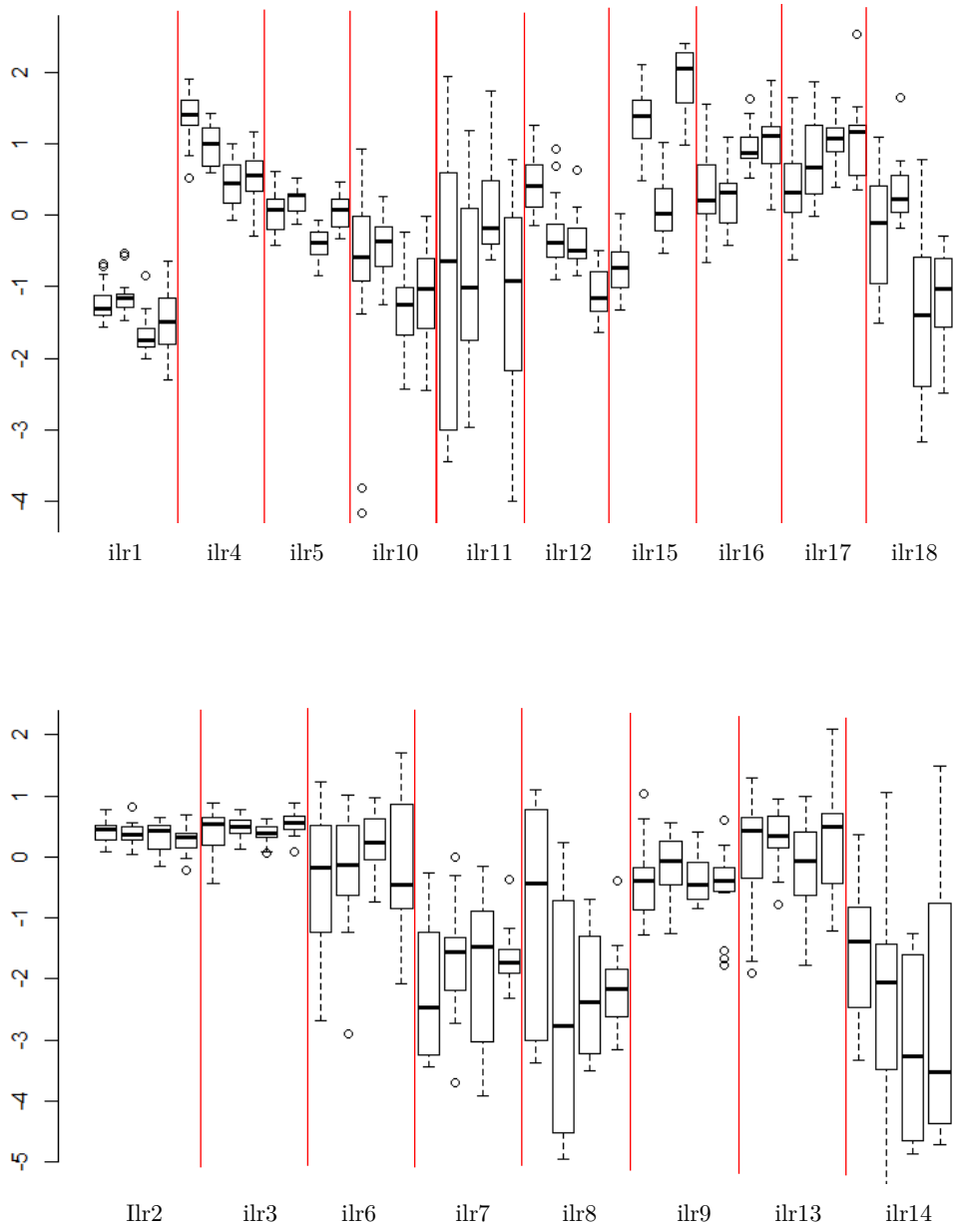


Figure 2 Boxplots of ilr coordinates [boxes left to right CS1, CS2, MR1, MR2] (top shows components that exhibit differences, bottom shows components with similar coordinates for sources or season

Psychological well-being: analysis and interpretation applying Compositional Data Analysis methods

M.Cortés[1], M.Sánchez[1], P.Galindo[1] and E.Jarauta-Bragulat[2]

[1] Universidad de Salamanca (USAL), Salamanca, Spain;
mariacortes@usal.es ; mersanbar@usal.es ; pgalindo@usal.es
[2] U. Politècnica de Catalunya - BarcelonaTech (UPC), Barcelona, Spain;
eusebi.jarauta@upc.edu

Summary

One of the applications of Psychology that makes use of Statistics is that which refers to the analysis of psychological well-being tests. However, this analysis was not systematic until Carol Riff proposed in 1989 a test to describe, analyze and interpret the psychological well-being of people. The model is based on six descriptive fields or dimensions of psychological well-being: self-acceptance, positive relations, autonomy, environmental mastery, purpose in life and personal growth. To measure these theoretical dimensions, an instrument known as "Scale of Physiological Well Being (SPWB)" was developed, with 120 items forming the original scale. Nowadays, there exist several different versions of the instrument in different languages and with different numbers of items. To interpret the results of the different dimensions, the scores of each of them must be added up and compared with the maximum and minimum possible score, since there are no existing ready-reckoners.

Application of the previous methodology generates certain problems when interpreting the results individually, since there is no normative or reference group with which to compare the results obtained. Applying methods of Compositional Data Analysis such as the centered log-ratio transformation (CLR), relative position ratios (RPR) of each individual can be obtained in relation to each of the indicators used. Positive values of the position ratio indicate that the subject is above the (geometric) mean of the normative group and negative values of the position ratio indicate that the subject is below the mean of the normative group. The average of the relative position ratios allows to obtain a global indicator or profile of subjective wellbeing for each individual, not only in relation to himself but also in the context of the group in which he has been analyzed.

Key words: Psychological well-being, psychometric, Compositional Data Analysis, Centered Logratio.

1. Introduction

Analysis and description of people well-being is a recurrent theme studied in the field of Psychology for a long time, and which is still under study today, due to its great interest. Well-being can be defined and interpreted from two perspectives: as a concept related to the

happiness of the individual and in this case we talk about the subjective well-being (SWB), or as a concept linked to the development of human potential and, in that case, it is called psychological well-being (PWB). This work focuses on this second approach, that is, the PWB is studied as a positive functioning of the individual in the different areas of his life.

In psychology, the measuring instruments are the psychological tests; the tests try to measure unobservable variables (psychological constructs) through observable indicators (items of the questionnaire) so that they can infer scores. In many occasions the psychological constructs are, as in this case, multidimensional.

Until 1989 studies on the welfare of people were carried out from a theoretical point of view and had not yet established a measurement instrument that had associated a measure of reliability or validity that would validate these studies. It is in this year when Carol Ryff proposes a multidimensional model for the study of psychological well-being; the model establishes six dimensions: positive relationships with other people (PRE), autonomy (AUT), control of the environment (CEN), personal growth (PGR), purpose in life (PLI) and self-acceptance (SAC).

The 1989 Carol Ryff PWB questionnaire was designed with 120 items, so that each dimension was formed by the sum of the scores on 20 items. Currently there are different versions of the questionnaire with a different number of items that have been validated in different countries. In the study that serves as an example for this work, the version of 14 items per dimension has been used, this is a total of 84 items, which is one of those recommended by the author. Therefore, to obtain the characterization and have a quantitative idea of the subjective psychological well-being of each person surveyed, the score of each of the 14 items of each dimension is carried out, resulting in a total of 6 scores.

The analysis and interpretation of the tests has traditionally been carried out through the analysis of the direct scores, the dimensions and the global construct. Thus, the higher the score in the "personal growth" dimension, it is inferred that the person puts more effort into developing their potential, in continuing to grow as a person and in maximizing their abilities. Depending on how high or low the individual's scores are in the sum of the different factors, a higher or lower level of psychological well-being can be established.

However, the way in which the dimensions are related to each other in a group is difficult to access if the data are not analysed as proportions in relation to the total instead of each value in isolation. Therefore, in this work we propose to take a step that complements the traditional study, making use of the concepts and methods of the Compositional Data Analysis. In this way, the focus of the study is not on the value of the degree of PBW of the individual, but on "how" or "in what proportions" that welfare is composed in such a way that the interrelations between the dimensions and the equilibria or imbalances between

them. On the other hand, the study can be expanded if subsets of individuals are studied in terms of sociodemographic variables.

2. Raw Data Analysis

The data set consists of information obtained through anonymous questionnaires online and in person from 623 people that answered the questionnaire completely. The confidentiality of the data has been explicitly guaranteed, since the answers to the test are not associated with personal data. On the other hand, three variable sociodemographic variables were made explicit that allow for partial studies that may be of interest; these variables are sex, educational level and age.

The statistical analysis of the data is done in a classical, when the questionnaires are not scaling (as it is our case) it can be done by calculating the statistics of each of the welfare dimensions. An illustration that is usual from this perspective is the one corresponding to values in Table 1 and represented in Figure 2.1, which shows the values of each dimension of two individuals selected at random (ID-387, ID-452), the maximum, the minimum and the average (arithmetic) of each dimension. The lines of values of both individuals would lead to affirm that they are quite similar in the first three and that in the last three ID-387 is slightly above ID-452; both individuals are located above the average in four of the dimensions (PRE, CEM, PLI, SAC).

Table 1. Statistical indicators for ID-387 and ID-452 (raw data).

	PRE	AUT	CEN	PGR	PLI	SAC
ID-387	77	59	71	73	71	78
ID-452	76	61	73	63	69	68
MIN	28	30	23	43	26	21
MAX	84	84	81	84	84	84
MEAN	65.02	61.37	57.12	66.68	63.46	61.50

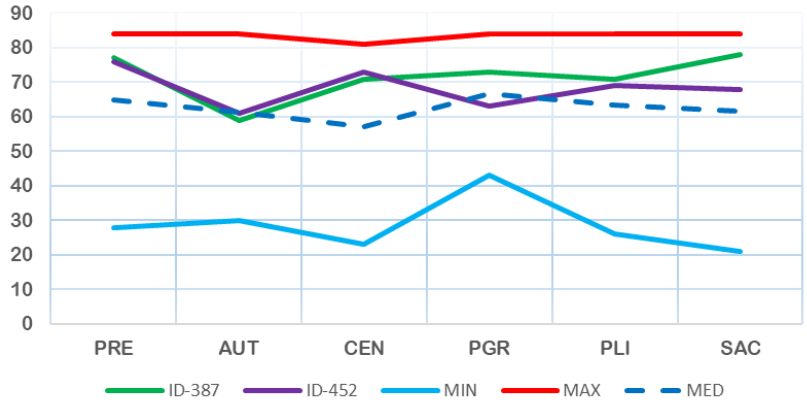


Figure 1. Graph of values of the direct data of two individuals (ID-387, ID-452) and the average, maximum and minimum statistical parameters.

3. The Compositional Data Approach

The compositional perspective starts calculating closed data from raw data values. Two "position ratios" can be obtained from corresponding proportions. First, applying the centered logratio transformation (CLR) for each individual, that is

(1)

The values obtained through Eq.(1) can be interpreted as a position ratio (PR1) of each of the dimensions of the individual in relation to their centrality estimator. This PR1 allows us to observe the balance or relative imbalance of each of the dimensions for each individual. For example, in Figure 2 the PR1 of two individuals are represented, together with three statistics: maximum, minimum and arithmetic mean of the values of the position ratios (values are in Table 2).

Table 2. Position ratios RP1 for ID-387 and ID-452, and statistical indicators.

	PRE	AUT	CEN	PGR	PLI	SAC
ID-387	-0.254	-0.463	-0.127	0.398	0.439	0.007
ID-452	0.077	-0.073	-0.166	0.064	0.114	-0.016
MIN	-0.508	-0.463	-0.601	-0.275	-0.441	-0.613
MAX	0.503	0.683	0.233	0.487	0.439	0.217
MEAN	0.041	-0.017	-0.091	0.072	0.017	-0.022

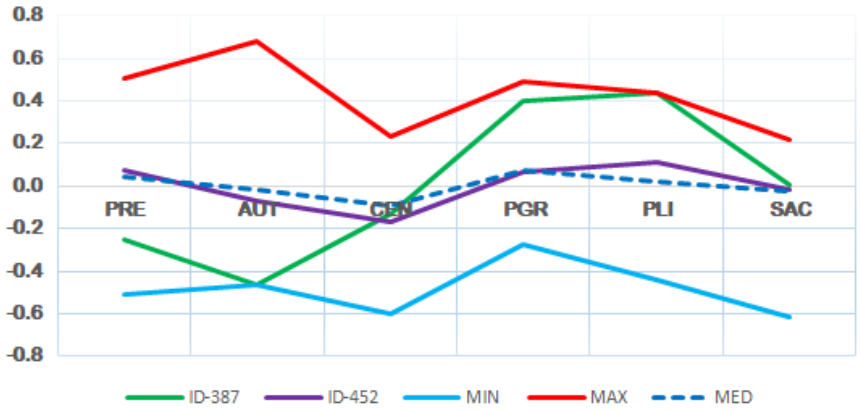


Figure 2. Graph of values of the position ratios PR1 of two individuals (ID-387, ID-452) and the lines corresponding to the values of the mean, the maximum and the minimum.

A second position ratio RP2 can be obtained to analyse the position of each dimension of the individual in relation to the average of each dimension. This position ratio is calculated by:

(2)

In Figure 3 the PR2 of the same two individuals are represented, together with three statistics: maximum, minimum and arithmetic mean of the values of the position ratios (values are in Table 3). Observe that, as it is obvious, arithmetic mean of that values is zero.

Note the remarkable difference between Fig.1 in relation to Fig.2 and Fig.3; both figures since, from this perspective, the individuals are noticeably different. In fact, ID-452 is a "balanced" individual in relative values with respect to the whole; on the other hand, the ID-387 is a more "unbalanced" individual and already has differences in its RP: the dimensions PRE, AUT are clearly below, the dimensions CEN, SAC are in the average and the dimensions PGR, PLI are clearly above of the set.

Table 3. Position ratios RP2 for ID-387 and ID-452, and statistical indicators.

	PRE	AUT	CEN	PGR	PLI	SAC
ID-387	-0.341	-0.493	-0.083	0.280	0.375	-0.018
ID-452	0.039	-0.054	-0.072	-0.005	0.099	0.009
MIN	-0.572	-0.493	-0.556	-0.358	-0.528	-0.652
MAX	0.422	0.630	0.271	0.353	0.375	0.232
MEAN	0.000	0.000	0.000	0.000	0.000	0.000

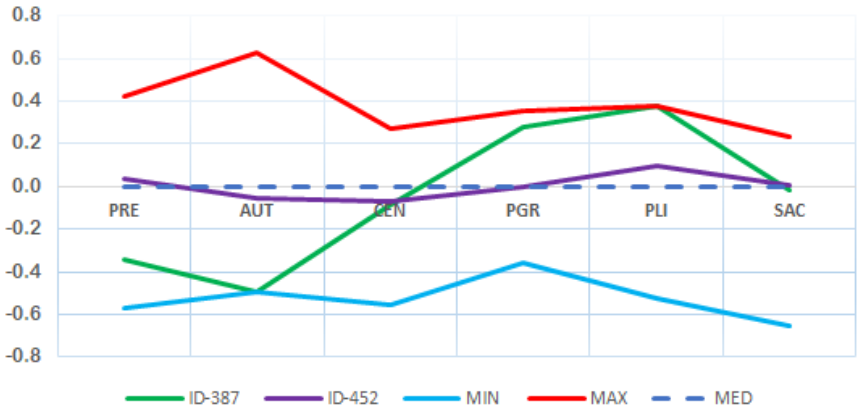


Figure 3. Graph of values of the position ratios PR2 of two individuals (ID-387, ID-452) and the lines corresponding to the values of the mean, the maximum and the minimum.

Conclusions

- The analysis of psychological well-being tests from a traditional perspective can be completed by applying the concepts and methods of Compositional Data Analysis.
- This new perspective allows analysing not only the dimensions by themselves but

also interrelated, being able to reveal elements that cannot be detected through traditional analysis.

- The centered logratio transformation (CLR) allows establishing for each individual a position ratio as a measure of the relative situation of each of the dimensions in relation to the set represented by its geometric mean.
- Additionally, for each individual, a second position ratio of each dimension can be obtained, calculating the logratio of each proportion between the geometric mean of all the proportions. This ratio indicates the relative position of each proportion of the individual in relation to the average value of all of them in the tested population.

Acknowledgements

The authors are grateful to:

(1) the Spanish Ministry of Economy and Competitiveness (MINECO), within the framework of the "CODA-RETOS / TRANSCODA" Project (Ref. MTM2015-65016-C2-1-R and MTM2015-65016-C2-2-R);

(2) the Agency for Management of University Grants and Research (AGAUR) of the Generalitat de Catalunya (GENCAT) within the framework of the project "Analysis of spatial and compositional data" (COSDA, Ref: 2014SGR551, 2014-2016).

References

Aitchison J (1986). The statistical analysis of compositional data. Monographs on statistics and applied probability. Chapman & Hall Ltd., London, p 416 (Reprinted in 2003 with additional material by The Blackburn Press).

Jarauta-Bragulat E, Colomer Y and Clotet R (2018). El sistema alimentario global: II aproximación cuantitativa al espacio agroalimentario de la Europa mediterránea. Revista Española de Estudios Agrosociales y Pesqueros, núm. 249, p. 15-38.

Ryff, C.D (1989). Happiness is everything, or is it? Explorations on the meaning of psychological well-being. Journal of Personality and Social Psychology, 57(6), 1069-1081.

Cortés-Rodríguez, M, Sanchez-Barba, M (2013). Biplot de Datos composicionales: una herramienta útil en el estudio de test psicológicos. (Trabajo Fin de Máster. Memoria conjunta). Universidad de Salamanca, España.

Corporate social responsibility and financial performance of Spanish hospitals. A compositional data approach with partial least squares

J. Creixans-Tenas[1], N. Arimany-Serrat[2], and G. Coenders[3]
[1]University of Vic, Spain
[2] University of Vic, Spain
[3]University of Girona, Spain; *germa.coenders@udg.edu*

Summary

In the context of the recent economic crisis, the financial situation of the Spanish health sector has been a major concern for responsible actors from both the public and private sphere, because of the decline of public spending and the increased demand due to population growth and ageing. The public health system seeks collaborative synergies with the private health system to achieve better health care results, reduce waiting lists and cope with financial pressure. The private health sector helps decongest the public system, and is a strategic collaborator, currently owing 57% of hospitals in Spain.

This paper analyses the financial statements of hospital companies in the Spanish private healthcare system using the Compositional Data (CoDa) methodology, and identifies the significant relationship between the financial log-ratios and corporate social responsibility (CSR) indicators. CSR practices tend to reduce turnover and increase leverage, while they have no impact on margin. From a methodological point of view, the paper has a number of firsts: the first application of pairwise log-ratios and of compositional prediction models in accounting and the first application of partial-least-squares structural equation models with CoDa. Our approach reduces the asymmetry, redundancy and outliers encountered in standard financial ratio analysis.

Key words: financial ratios, Compositional Data analysis (CoDa), pairwise log-ratios, partial least squares structural equation models, corporate social responsibility.

1 Introduction

In the context of the recent economic crisis, the financial situation of the Spanish health sector has been a major concern for responsible actors from both the public and private sphere. According to the Report on the Situation of Private Health (Institute for the Development and Integration of Health –IDIS), the private health sector decongests the public system, and is a strategic collaborator for better healthcare. The private health sector currently has 452 hospitals in Spain (57% of country total). The private hospital sector represents 58% of hospital groups, 39% of independent hospitals, and 3% of insurance companies.

In the current economic environment, business profitability is one of the aspects of greatest interest for private hospital companies (Creixans-Tenas and Arimany-Serrat, 2018). In this respect, non-financial variables have received increasing attention because of their contribution to improving profitability. This includes Corporate Social Responsibility (CSR) practices, which have already proved their worth in the hospital sector (Creixans-Tenas and Arimany-Serrat, 2018; Vélez-González et al., 2011). The aim of this

paper is to relate CSR practices in private Spanish hospitals to their profitability, as revealed by their financial ratios.

2 Method

2.1 Financial ratios as carriers of relative information about positive magnitudes

Financial ratios, i.e., ratios comparing the magnitudes of accounts in financial statements, constitute a case of researchers' and professionals' interest in relative rather than absolute account magnitudes. The relative nature of financial ratios enables them to evaluate the company's position compared to its counterparts in the industry or to itself along time, taking into account differences in firm size. Since the beginning of the last century financial ratios have been used both in practical management performance and strategic assessment, and in research relating them to other financial or non-financial variables (Barnes, 1987; Willer do Prado et al., 2016).

When treated as variables in statistical analyses, financial ratios have been reported to have a number of serious statistical and practical problems, including asymmetry, outliers, redundancy, severe non-normality, and even dependence of the results on the arbitrary decision regarding which account appears in the numerator and which in the denominator (see Linares-Mustarós et al., 2018 for a review).

Compositional Data Analysis (CoDa) is a standard methodology for analysing the relative importance of magnitudes (van den Boogaart and Tolosana-Delgado, 2013; Pawlowsky-Glahn et al., 2015). CoDa treats magnitudes (i.e., account values) in a symmetric fashion in such a way that results do not depend on numerator and denominator permutation, tends to reduce outliers and non-normality, and treats redundancy by acknowledging the fact that no analysis will require more variables than there are account magnitudes to be compared.

After the seminal work of Aitchison (1986) over thirty years of development have led to a well-established standard CoDa toolbox which is covered in text books (van den Boogaart and Tolosana-Delgado, 2013; Filzmoser et al., 2018; Greenacre, 2018a; Pawlowsky-Glahn et al., 2015). Recently, CoDa has also been applied in finance to answer research questions concerning relative magnitudes. Examples include crowdfunding (Davis et al., 2017), financial markets (Ortells et al., 2016), municipal budgeting (Voltes-Dorta et al., 2014), investment portfolios (Belles-Sampera et al., 2016; Boonen et al., 2019; Glassman and Riddick, 1996), product portfolios (Joueid and Coenders, 2018) and insurance pricing (Verbelen et al., 2018). Within the accounting field, CoDa has already been successfully applied with the purpose of clustering firms with similar financial statement structures (Linares-Mustarós et al., 2018). To the best of our knowledge this paper is the first to use CoDa to predict financial statement structure from non-financial variables.

2.2 Financial statement accounts as compositional data

Compositional Data are positive vector variables carrying information about the relative size of their D components to one another (Aitchison, 1986):

$$\mathbf{x} = \left(x_1, x_2, \ldots, x_D\right) \text{ with } x_j > 0,\ j = 1,2,\ldots,D$$

Some rules have to be followed in order to introduce financial accounts in a D-part composition, which boil down to avoiding negative accounts and account overlap.

 ISBN 978-84-947240-2-2

Even if sometimes financial ratios involve accounts which may be negative, its use is advised against in the financial literature, because they can cause a discontinuity, outliers, or even a reversal of interpretation when the account which may be negative is in the denominator (Lev and Sunder, 1979). Negative accounts are also advised against from the point of view of measurement theory. Computing a ratio is a meaningful operation only for variables in a ratio scale, which need to have a meaningful absolute zero (Stevens, 1946) and thus no negative values.

In general, accounts are negative because they subtract of other positive accounts, which are the ones to be used. This means, for instance, that one should directly use revenues and costs rather than profit. This limitation implies no loss of information whatsoever. A ratio conveying the same information as the classical margin ratio (*profit/revenues*) can be constructed from only the non-negative magnitudes of revenues and costs. Let x_1=revenues, x_2=costs, $x_3=x_1-x_2$=profit. The always positive revenues over costs ratio (x_1/x_2) can easily be shown to be just a transformation of the problematic profit over revenues ratio (x_3/x_1):

$$\frac{x_1}{x_2}=\frac{x_1}{x_1-x_3}=\frac{1}{\frac{x_1-x_3}{x_1}}=\frac{1}{1-\frac{x_3}{x_1}}$$

It must also be taken into account that components may not overlap. For instance, one could not use x_4:assets and x_5:fixed assets because x_5 is a part of x_4. In compositional data terminology, x_4:assets is an amalgamation of x_5:fixed assets and x_6:current assets. Using both amalgamations and their constituent parts is extremely problematic (Pawlowsky-Glahn et al., 2015). Rather, the choice between using only the amalgamation or only the individual parts should be made at the problem definition stage and cannot be changed afterwards (van den Boogaart and Tolosana-Delgado, 2013). It is not essential to use all constituent parts, which is referred to as a subcomposition in compositional data terminology. Accordingly, the feasible choices to handle x_4 to x_6 are: a) to use only x_4; b) to use x_5 and x_6; c) to use only x_5; and d) to use only x_6.

In this paper, the components are D=4 positive and non-overlapping financial accounts of private hospital companies, namely x_1: revenues; x_2: costs; x_3: liabilities; and x_4: assets. Accounts were obtained from the SABI (Iberian Balance sheet Analysis System) database, developed by INFORMA D&B in collaboration with Bureau Van Dijk. Search criteria were private hospital companies in Spain with available data for 2016 (n=107).

These accounts are very relevant because they make it possible to compute some of the most common profitability, turnover, margin and leverage ratios. According to the DuPont analysis method, the return on shareholders' equity (ROE, the key measure of profitability) can be decomposed as the product of turnover, margin and leverage, according to the following financial ratios, some of which are computed from magnitudes which may be negative:

$$\text{ROE}=\text{profit/net worth}=(x_1-x_2)/(x_4-x_3)$$
$$\text{Turnover}=\text{revenues/assets}=x_1/x_4$$
$$\text{Margin}=\text{profit/revenues}=(x_1-x_2)/x_1$$
$$\text{Leverage}=\text{assets/net worth}=x_4/(x_4-x_3)$$

2.3 Transformations. Pairwise financial log-ratios

The usual approach to CoDa is using standard statistical methods on transformed data. *Logarithms of ratios* are the commonest transformation in CoDa (Pawlowsky-Glahn et al., 2015). A log-ratio involving

only two components is computed as:

$$y = \log_2\left(\frac{x_1}{x_2}\right),$$

where $\log_2$ stands for the logarithm to base 2. This base is recommended by Müller et al. (2018) on the basis of interpretability. Unlike a standard ratio, which is bounded between zero and infinity, a log-ratio is symmetric in the sense that its range is from minus infinity to plus infinity. Besides, permuting the numerator and denominator affects no other property of the log-ratio than the sign. Furthermore, if one of the components being compared is close to zero, it may lead to an outlying standard ratio when placed in the denominator and to a typical standard ratio when placed in the numerator. For log-ratios placement makes no difference.

It can be show that just D–1 log-ratios contain all information about the relative importance of D components, thus preventing the redundancy problems encountered in the financial literature when using a very large number of ratios, some of which are unavoidably exact functions of other ratios (Chen and Shimerda 1981).

In the words of Barnes (1987, p. 456) there is a need to "identify those ratios which contain complete information about a firm while minimising duplication". Several choices are possible to select D–1 log-ratios (Egozcue et al., 2003), and the researcher enjoys some freedom in tailoring log-ratios to their interpretability and to the research questions involved (Greenacre, 2018a; 2018b; Joueid and Coenders, 2018). Greenacre (2018b) recommends computing only ratios among pairs of components on the grounds of their simpler interpretation, which he appropriately terms *pairwise log-ratios*. In order to avoid redundancy, Greenacre (2018b) recommends drawing a graph in which the parts are vertices (nodes) and the log-ratios are connections (edges). The graph must necessarily be connected (all parts have to participate in at least one log-ratio) and acyclic (there may not be closed circuits, that is, when following the edges of the graph from one vertex to any other vertex, no vertex can be visited twice).

While any graph fulfilling these conditions will do the job, statistically speaking, and even automatic selection methods can be used, it is good practice to use a graph with substantive interpretation, based on expert knowledge (Greenacre, 2018b) or on the research purpose. In our case, the research purpose is DuPont analysis and we thus look for ratios conveying the notions of margin, turnover and leverage. The graph is in Figure 1. By definition, turnover compares revenues and assets:

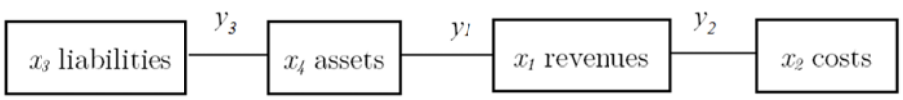
$$y_1 = \log_2\left(\frac{x_1}{x_4}\right)$$

As argued above, comparing revenues and costs provides a notion of margin:

$$y_2 = \log_2\left(\frac{x_1}{x_2}\right)$$

Finally, comparing liabilities and assets provides a notion of leverage:

$$y_3 = \log_2\left(\frac{x_3}{x_4}\right)$$



 ISBN 978-84-947240-2-2

Figure 1: Graph diagram used in this paper. Each edge between two nodes represents their log-ratio.

Any other log-ratio is unnecessary as it results from the previous three. For instance, if one would like to compute a cost-to-asset log-ratio:

$$y_4 = \log_2\left(\frac{x_2}{x_4}\right) = \log_2\left(\frac{x_1}{x_4}\right) - \log_2\left(\frac{x_1}{x_2}\right) = y_1 - y_2 \cdot$$

2.4 Modelling

Financial ratios are commonly related to non-financial variables. Once the y_1 to y_3 log-ratios have been computed, they can be introduced as either dependent or explanatory variables in statistical models, ranging from linear models such as MANOVA and ordinary least squares regression (Tolosana-Delgado and van den Boogaart, 2011) to structural equation models –SEM– (Kogovšek et al., 2013), which include covariance-based SEM and partial-least-squares SEM (PLS-SEM). The models can then be estimated with standard methods and software. When some variable(s) are measured with multiple indicators, as CSR practices in our case, SEM are called for. The model specification is as follows:

y_1 to y_3 are dependent variables, each being its own dimension, i.e. single item measures. They all depend on CSR in the inner model. d_1 to d_5 are dummy-coded indicators of the CSR dimension indicating if the web site shows CSR accreditations:

d_1: Global Reporting Initiative (GRI, 35% of sample hospitals).

d_2: EFQM quality accreditation (39%).

d_3: JCI quality accreditation (26%).

d_4: ISO 50001 accreditation (20%).

d_5: ISO 26000 accreditation (15%).

In the outer model relating the indicators to CSR, we conceptualise d_1 to d_5 as formative indicators because socially responsible firms may be so by adopting them in any combination. Thus, correlations among d_1 to d_5 are not expected to be high and should actually not be high, for the sake of collinearity (Diamantopoulos and Winklhofer, 2001). The path diagram is shown in Figure 2.

There are two main arguments for using PLS-SEM rather than covariance-based SEM. First, PLS-SEM is a more convenient and flexible approach to formative indicators (Hair et al., 2011; 2012; Lee et al., 2011). Second, PLS-SEM is suitable for small sample sizes, although when the sample size is small, the bootstrap procedure used to test the model is no-longer robust to severe non-normality (Hair et al., 2012). To the best of our knowledge, this paper is the first compositional application of PLS-SEM.

3 Results

Anderson-Darling (A-D) normality tests were carried out for y_1 to y_3 (adtest command in the R package robCompositions, Table 1). As argued above, the lack of extreme non-normality is especially important for small sample sizes. y_1 to y_3 can indeed be considered to be non-normal following the A-D tests, but the relatively low skewness and kurtosis would put the analysis on the safe side. For comparison, the same normality diagnostics are presented for the standard financial ratios in DuPont analysis, whose use would be unwise for PLS-SEM with this sample size due to very extreme skewness and high positive kurtosis which indicates the presence of outliers.

 ISBN 978-84-947240-2-2

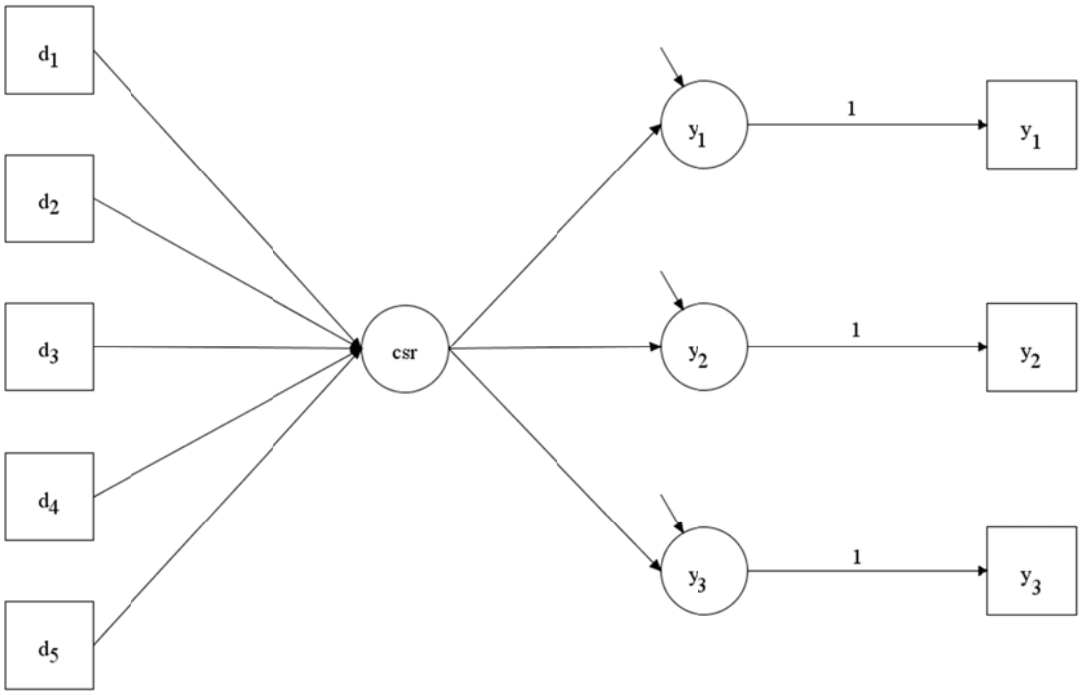


Figure 2: SEM path diagram with outer model measuring CSR (left) and inner model explaining pairwise log-ratios from CSR (right).

Table 1: Normality assessment of pairwise log-ratios and selected standard financial ratios.

	Skewness	Kurtosis	A-D test p-value
$y_1=\log_2(x_1/x_4)$	-0.14	0.19	0.927
$y_2=\log_2(x_1/x_2)$	1.09	2.57	<0.001
$y_3=\log_2(x_3/x_4)$	-1.03	1.24	<0.001
ROE$=(x_1-x_2)/(x_4-x_3)$	0.38	25.44	<0.001
Turnover$=x_1/x_4$	1.56	2.91	<0.001
Margin$=(x_1-x_2)/x_1$	0.67	1.70	0.004
Leverage$=x_4/(x_4-x_3)$	6.35	50.71	<0.001

PLS estimation was carried out with the R package plspm.formula. The centroid weighting scheme was used, with 20,000 replications for the bootstrap procedure. As recommended by Hair et al. (2011) and Ruiz-Molina et al. (2018), the outer formative model was assessed from weight sign and significance, standardized loading sign, size and significance, variance inflation factor (VIF) and relevance from a content validity point of view. According to all these criteria we decided to keep all CSR practices in the model.

Table 2 shows the standardized model estimates, including weights and loadings and their

 ISBN 978-84-947240-2-2

corresponding bootstrapped confidence intervals, VIF, and R squared values. According to the inner model paths, CSR practices tend to reduce turnover and increase leverage, while they have no significant impact on margin. A cautionary explanation would be the increased assets needed to implement some CSR practices, and the consequent need for additional financial resources (i.e. liabilities).

Table 2: Standardized PLS estimates. Bootstrap means, standard errors, 95% confidence intervals, VIF and R-squares.

Outer weights	Mean	Std.Error	perc.025	perc.975	VIF
CSR–d_1	0.671	0.866	-1.069	2.29	2.216
CSR–d_2	0.220	0.722	-1.225	1.59	1.693
CSR–d_3	0.236	0.916	-1.523	1.97	1.744
CSR–d_4	0.468	1.086	-1.682	2.41	1.711
CSR–d_5	0.854	1.093	-1.310	2.85	1.696
Outer loadings	Mean	Std.Error	perc.025	perc.975	
CSR–d_1	0.700	0.189	0.247	0.956	
CSR–d_2	0.557	0.228	0.028	0.899	
CSR–d_3	0.518	0.249	-0.014	0.912	
CSR–d_4	0.574	0.249	0.010	0.934	
CSR–d_5	0.660	0.213	0.177	0.959	
Inner paths	Mean	Std.Error	perc.025	perc.975	R-Squares
CSR–y_1	-0.242	0.091	-0.404	-0.067	0.067
CSR–y_2	-0.049	0.156	-0.284	0.254	0.027
CSR–y_3	0.215	0.103	0.006	0.391	0.057

4 Concluding remarks

Compositional data analysis offers distinct advantages in statistical modelling of financial statements, by reducing redundancy, asymmetry, non-normality and outliers. Once log-ratios have been computed, statistical analysis becomes standard in all respects, and researchers may use their favourite models, estimation methods and software. The proposed method of financial statement analysis boils down to computing logarithms of financial ratios between pairs of the D financial accounts of interest in such a way that all accounts are connected to some other. Once D non-overlapping positive accounts of interest have been selected, D–1 log-ratios suffice. CoDa can be understood as a manner to select a minimum number of variables which carry all information about the relative importance of any account to any other. Log-ratios also tend to be less deviant from the normal distribution and to contain fewer outliers.

It must be taken into account that pairwise log-ratios are not isometric, in other words, they do not preserve distances among firms (Egozcue et al., 2003). They are thus unfit for statistical analyses based on distances such as cluster analysis. The use of isometric log-ratios in financial statement analysis is discussed in Linares-Mustarós et al. (2018). Isometric log-ratios are perfectly appropriate for the PLS-SEM analysis we carried out in this paper. We use pairwise log-ratios only because of their increased simplicity and interpretability in terms of the common practices in financial statement analysis. It must also be taken into account that PLS-SEM is not generally affine equivariant, so that, when using pairwise log-ratios, results depend on the chosen graph diagram. PLS-SEM is not generally orthogonal

equivariant either, so that, when using isometric log-ratios, results also depend on the orthonormal basis chosen (see Filzmoser et al., 2018 for a general discussion on equivariance). This has two consequences. On the one hand, results are interpretable in as much as log-ratios are interpretable in themselves. On the other hand there seems to be no obvious gain in using isometric log-ratios, so that the log-ratios may be chosen on the basis of interpretability alone.

References

Aitchison, J. (1986). *The statistical analysis of compositional data. Monographs on statistics and applied probability*. London: Chapman and Hall.

Barnes, P. (1987). The analysis and use of financial ratios: A review article. *Journal of Business Finance & Accounting 14*(4), pp. 449–461.

Belles-Sampera, J., M. Guillen and M. Santolino (2016). Compositional methods applied to capital allocation problems. *Journal of Risk 19*(2), pp. 15–30.

Boonen, T., M. Guillén and M. Santolino (2019). Forecasting compositional risk allocations. *Insurance Mathematics and Economics 84*, pp. 79–86.

Van den Boogaart, K.G. and R. Tolosana-Delgado (2013). *Analyzing compositional data with R*. Berlin: Springer.

Chen, K.H. and T.A. Shimerda (1981). An empirical analysis of useful financial ratios. *Financial Management 10*(1), pp. 51–60.

Creixans-Tenas, J. and N. Arimany-Serrat, N. (2018). Influential variables on the profitability of hospital companies. *Intangible Capital 14*(1), pp. 171–185.

Davis, B.C., K.M. Hmieleski, J.W. Webb and J.E. Coombs (2017). Funders' positive affective reactions to entrepreneurs' crowdfunding pitches: The influence of perceived product creativity and entrepreneurial passion. *Journal of Business Venturing 32*(1), pp. 90–106.

Diamantopoulos, A. and H.M. Winklhofer (2001). Index construction with formative indicators: An alternative to scale development. *Journal of Marketing Research 38*(2), pp. 269–277.

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), pp. 279–300.

Filzmoser, P., K. Hron and M. Templ (2018). *Applied compositional data analysis with worked examples in R*. New York: Springer.

Glassman, D.A. and L.A. Riddick (1996). Why empirical international portfolio models fail: evidence that model misspecification creates home asset bias. *Journal of International Money and Finance 15*(2), pp. 275–312.

Greenacre, M. (2018a). *Compositional data analysis in practice*. New York: CRC press.

Greenacre, M. (2018b). Variable selection in compositional data analysis using pairwise logratios. *Mathematical Geosciences*. Doi: 10.1007/s11004-018-9754-x

Hair, J.F., C.M. Ringle and M. Sarstedt (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing theory and Practice 19*(2), pp. 139–152.

Hair, J.F., M. Sarstedt, T.M. Pieper and C.M. Ringle (2012). The use of partial least squares structural equation modeling in strategic management research: a review of past practices and recommendations for future applications. *Long Range Planning 45*(5–6), pp. 320–340.

Joueid, A. and G. Coenders (2018). Marketing innovation and new product portfolios. A compositional approach. *Journal of Open Innovation: Technology, Market and Complexity 4*, 19.

Kogovšek, T., G. Coenders and V. Hlebec (2013). Predictors and outcomes of social network compositions. A compositional structural equation modeling approach. *Social Networks 35*(1), pp. 1–10.

Lee, L., S. Petter, D. Fayard and S. Robinson (2011). On the use of partial least squares path modeling in accounting research. *International Journal of Accounting Information Systems 12*(4), pp. 305–328.

Lev, B. and S. Sunder (1979). Methodological issues in the use of financial ratios. *Journal of Accounting and Economics 1*(3), pp. 187–210.

Linares-Mustarós, S., G. Coenders and M. Vives-Mestres (2018). Financial performance and distress profiles. From classification according to financial ratios to compositional classification. *Advances in Accounting 40*, pp. 1–10.

Müller, I., K. Hron, E. Fišerová, J. Šmahaj, P. Cakirpaloglu and J. Vančáková (2018). Interpretation of compositional regression with application to time budget analysis. *Austrian Journal of Statistics 47*(2), pp. 3–19.

Ortells, R., J.J. Egozcue, M.I. Ortego and A. Garola (2016). Relationship between popularity of key words in the Google browser and the evolution of worldwide financial indices. In J.A. Martín-Fernández and S. Thió-Henestrosa (Eds.), *Compositional data analysis. Springer proceedings in mathematics & statistics, Vol. 187,* pp. 145–166. Cham: Springer.

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

Ruiz-Molina, M.E., D. Servera-Francés, F. Arteaga-Moreno and I. Gil-Saura (2018). Development and validation of a formative scale of technological advancement in hotels from the guest perspective. *Journal of Hospitality and Tourism Technology 9*(3), pp. 280–294.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science 103*, pp. 677–680.

Tolosana-Delgado, R. and K.G. van den Boogaart (2011). Linear models with compositions in R. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional data analysis. Theory and applications,* pp. 356–371. New York: Wiley.

Vélez-González, H., R. Pradhan and R. Weech-Maldonado, R. (2011). The role of non-financial performance measures in predicting hospital financial performance: The case of for-profit system hospitals. *Journal of Health Care Finance 38*(2), pp. 12–24.

Voltes-Dorta, A., J.L. Jiménez and A. Suárez-Alemán (2014). An initial investigation into the impact of tourism on local budgets: A comparative analysis of Spanish municipalities. *Tourism Management 45*, pp. 124–133.

Verbelen, R., K. Antonio and G. Claeskens (2018). Unravelling the predictive power of telematics data in car insurance pricing. *Journal of the Royal Statistical Society Series C 67*(5), pp. 1275–1304.

Willer do Prado, J., V. de Castro Alcântara, F. de Melo Carvalho, K. Carvalho Vieira, L.K. Cruz Machado and D Flávio Tonelli (2016). Multivariate analysis of credit risk and bankruptcy research data: a bibliometric study involving different knowledge fields (1968–2014). *Scientometrics 106*(3), pp. 1007–1029.

Compositional analysis approach in the measurement of social-spatial segregation trends. A case study of Guadalajara, Jalisco, Mexico.

M.A. Cruz[1], M.I. Ortego[1], and E. Roca[1]

[1]Department of Civil and Environmental Engineering.
Universitat Politècnica de Catalunya BarcelonaTech, Spain;
marco.antonio.cruz@upc.edu

Abstract

The place in which we live affects our outlook on life; it is through the place of residence and its surroundings where our relationships, thoughts and opportunities are born and shaped. Different authors have highlighted the internal existing social differences in cities as a consequence of the neoliberal system in which we live. The mercantile logic that affects urban spaces incentives the dichotomy winners-losers in the current urban landscape and leads to the differentiation and unequal distribution of certain social groups within the urban space. This clear differentiation in distribution of social groups in the urban space has been called socio-spatial segregation. This concept arises from the urban sociology, the first studies were focused on the differentiation of ethnicity and income level to identify the most vulnerable groups and of mitigate their current situation through different policies.

A more significant number of variables belonging to different dimensions (social, economic, political and environmental) have been incorporated into the study of this phenomenon, traditionally addressed by different disciplines such as sociology, geography and anthropology. Nonetheless, few studies have addressed it from a multivariate analysis approach. Moreover, the few existing studies with a multivariate statistical analysis ignored or did not know the compositional nature of their data.

The objective of the present study is to introduce the compositional data analysis in urban studies to better understand socio-spatial segregation in the different urban contexts. Specifically, the analysis of social-spatial segregation considering the compositional nature of the data in the city of Guadalajara, Mexico, is carried out. Socio-economic variables from census data of approximately 13,520 urban blocks grouped in 395 colonias and seven urban districts are used to carry out this study through the most straightforward compositions of two components. Additionally, principal component analysis and cluster analysis are performed to identify the socio-economic distribution within the territory. The analysis is complemented with the use of geographic information systems (GIS) at different urban scales.

Based on Aitchison log ratio approach, the results are consistent with the segregation processes that date back to the foundation of the city. Through cluster analysis and principal component analysis, an evident polarization between the Minerva district and the rest of the areas is shown.

Key words: Compositional analysis, segregation, Guadalajara.

1 Introduction

In its broader conception, the existence of differentiation or unequal distribution of certain social groups within the urban space is known as urban segregation (Brun 1994). The term segregation emerged from urban sociology, and its study is traced back to the first half of the 20th century by the Chicago School of Sociology (Dawkins, Reibel and Wong 2007). From its origins, the social division of space was strictly linked to the concepts of income and race. However, nowadays when talking about segregation it is necessary to broaden the spectrum of elements that generate it and treat it as a dynamic process caused by different factors with different intensities that can accelerate or sustain the phenomena of segregation (Donat 2018); (Massey and Denton 1988); (Peach 1996); (Subirats 2004).

Although the study of segregation began around 1920, the first measurement is attributed to Jahn in 1947 with the index of dissimilarity (Jahn, Schmid and Schrag 1947). In his work, the author highlighted the applicability of his index, which can be applied to any population or class. However, according to Cowgill and Cowgill (1951), this index had two major flaws. The first of them, its difficulty to calculate it and the second, a lack of precision in the measurement differences in concentration and dispersion.

The interest of different authors for this phenomenon, coupled with technological advances, led to the appearance of a large number of indexes. For this reason, Massey and Denton (1988) emphasized on the existing state of theoretical-methodological disorder and a minor consensus in the use of indicators by the researchers of that time. In their study, the twenty most relevant existing segregation indexes in the literature were selected and classified in five dimensions: uniformity, exposure, concentration, centralization and clustering.

Concerning the first dimension, Massey and Denton (1988) proposed the dissimilarity index of Duncan and Duncan (1955), which measures the degree to which the groups are distributed differently in the urban space. The Lieberson index (1981) with respect to the exposure dimension was selected to represent the degree to which the members of different groups share common areas. The index of relative concentration was proposed to measure the degree of agglomeration of a group in the urban space within the dimension of concentration. On the other hand, the absolute centralization index was chosen to evaluate the degree to which the groups are located in the center of the cities. Finally, the spatial proximity index was selected to measure the proximity between groups in the urban space within the dimension of clustering.

The lack of the element of spatiality in the measurement of segregation represented the greatest criticism for the former indexes (Romero Mares and Hernández Lozano 2015); (Johnston, Poulsen and Forrest 2015); (Wong 2015); (Lloyd, Catney and G.Shuttleworth 2015). Likewise, the proposed indexes in their different dimensions simplified the measurement of segregation in considering an average value applied to race and religion as a total indicator of segregation of a territory without showing the various nuances that may exist within it (Johnston, Poulsen and Forrest 2015).

The ambiguity of the concept of segregation (Duncan and Duncan 1955); (Romero Mares and Hernández Lozano 2015); (Donat 2018), together with the incompatibility in the census information (temporality, statistical geographical areas and indicators), has made impossible the evaluation and comparison between cities and population groups in the different national contexts (Kertzer and Arel 2002); (Mateos 2015).

Over time, different methods for measuring segregation have been proposed. These methods have evolved based on the disaggregation of census information, statistics, the dimensions of study and the available technologies for the treatment of data (Lloyd, Shuttleworth and Wong 2015); (Romero Mares and Hernández Lozano 2015). Nonetheless, few studies have addressed it from a statistical multivariate analysis approach (Schteingart 2015). Moreover, the few existing studies with a multivariate statistical analysis ignored or did not know the compositional nature of their data such as the ones elaborated by CONAPO (2010), Romero and Hernández (2015), Shteingart (2015) and Jiménez and Donat (2018) among others.

For such a motive, the objective of this study is to introduce compositional data analysis in urban studies to better understand socio-spatial segregation in the different urban contexts. Specifically, this approach is applied to Guadalajara, Mexico. For this purpose, census information of approximately 13,520 urban blocks grouped into 395 colonias and seven urban districts of the city is analyzed as two-part compositions.

2 Case study of Guadalajara

Recent urban policies characterized by the hegemony of certain economic and political forces have made of Guadalajara a place where the interests of class, the logic of wealth and accumulation have overlapped rational urban planning. As mentioned by Kempen (2002) , cities are not naturally divided, its division is the product of an intentional and active act of those that have the power to do it. As regards to Guadalajara the division of space dates back from its foundation in 1542. The Spanish Crown, through its urban ordinances and know-how, established a defined geometric scheme within the urban fabric and a clear social hierarchy in the city (López Moreno 2001).

Among the criteria of the know-how by the Spanish Crown, in order to guarantee access to water, cities should be located in the proximity of a river (Vázquez 1989). In the latter sense, the city of Guadalajara settled on to one side of the San Juan de Dios River. A natural border that internally divided the city (Aceves, Torre and Safa 2004). The local bourgeoisie, the rich and renowned people, would concentrate on the west side of the river (except for the 200 indigenous allies who were responsible for the defense of the city). On the other hand, the indigenous population with no nobility titles was located east of the river, installing a clear division of the urban space, a Spanish city versus the city of Indians (Vázquez, 1989).

Even though in 1896 the San Juan de Dios River was forced underground and the Independencia Causeway was built over it, the urban planned growth continued on the west side of the river, such as the Colonias in the period 1894-1924. Product of foreign

capital and in its beginnings inhabited by foreigners and wealthy families, Colonias were homogeneous subdivisions that responded to commercial interests that sought the increase of the value of the land and that accentuated the east-west and poor-rich dichotomy of the city (Alvizo 2013); (Barajas and Muñoz 2006); (Cabrales and Canosa 2001); (Doñán 2013); (López Moreno 2001); (Rivera 2012); (Vázquez 1989).

In addition, the incorporation of neoliberal policies and the modification of structural reforms in Mexico in the period from 1982 to 1988, provoked the abandonment of the State in its role of urban planner what has meant a more significant social division and differentiation of the space in the Mexican cities (Marrufo and Bass Zavala 2015); (Moreno Pérez 2015); (Rivera 2013).

3 Methodology

To carry out the present study, a descriptive multivariate statistical analysis is performed, the compositional nature of the data and their properties are considered (Aitchison 1986); (Pawlowsky-Glahn and Egozcue 2006). This approach implies that the data used is in a restricted space called the Simplex; likewise , the data used is part of a whole, is positive, is in a range between zero and a positive number, and is subjected to the sum of a constant (Pawlowsky-Glahn and Egozcue 2006). This fact conditions the relationship that variables have to one another. Data does not vary independently as they would if they were not subject to the sum of the constant and that can be seen in the variance-covariance structure. The constant sum constraint forces at least one covariance to be negative and at least con correlation between elements will be negative. The latter means that coefficients between elements range between -1 and +1, which leads to the existence of spurious correlations.

To overcome the consequences of working with compositional data and to validate the statistical analysis, the log-ratio approach proposed by Aitchison (1986) is applied. This approach allows us to work with log ratios of compositions as if they were real random variables, and therefore the multivariate classic statistics tools can be applied. Moreover, the analysis is based on the relative information between the components and not on their absolute values (Aitchison 1986).

The indicators used in this study are simple two-part compositions, which are grouped and analyzed in different dimensions (i.e., social and economic), see Equation (1). As a result of having the sum-constraint the data is transformed with the log-ratio approach. Hence, the second component analyzed will be 100 minus the sum of the first component, see Equation (2).

$$X = C[x_1, x_2] \in S^2 \quad (1) \qquad log(X) = \frac{x_1}{1-x_1} \quad (2)$$

Where:

X = Compositional vector of two parts.

C = Stands for closure. The vector has been rescaled making the components add to 100.

x_1 , x_2 = Parts of our compositional vector.

S^2 = Subset of D=2 dimensional real space in the simplex.

Important care should be addressed in the values used in the numerator and the denominator of Equation (2). The above has a direct influence on the results obtained. Therefore, in this study, the aspects of the compositions that were considered as positive aspects were placed in the numerator and negative elements in the denominator (e.g., log ratio of educated population and uneducated population).

Once the information is presented in log ratios, a descriptive multivariate statistical analysis is carried out in R. First, a Principal Component Analysis (PCA) is done for a dimensionality reduction. This analysis explains the variability of our data through a set of new dimensions. PCA illustrates the behavior of the observations, the relationship and the direction of the variables. Moreover, it helps to identify clusters of observations in the data. Second, a hierarchical cluster analysis using Wards method and Euclidean distances is performed. Cluster analysis allows to group observations with similar behavior and allows to know the characteristics of the observations grouped in the different groups.

Finally, Geographic Information Systems (GIS) were used to incorporate the spatial element into the study. These systems allow the visualization of the distribution and the social division of the space based on the different indicators and dimensions analyzed.

4 Data

The present study includes the analysis of census information of approximately 13 520 urban blocks, which are grouped in 395 colonias of the city and which in turn are grouped into seven large urban districts. Therefore, this study involves the use of different sources of information for its realization and which are described below.

4.1 Geospatial vector information

The vector information corresponding to the geographic information systems in its shapefile format is obtained from two sources. The data from the 13,520 urban blocks and the territorial limits of the city of Guadalajara is obtained from the National Institute of Statistics and Geography (2010). On the other hand, the territorial delimitation of the 395 colonias and the seven urban districts is obtained from GeoGDL (2019).

4.2 Census information

The census information in Mexico is carried out every ten years. For this reason, the most recent census information at the urban block level corresponds to the year 2010. Information corresponding to the indicators used in this study is obtained from the 2010

Population and Housing Census (INEGI 2010). In this information, unique identification codes are presented for the different urban blocks. Therefore, the data is linked with Geographic Information Systems, which facilitates the processing of data at different scales such as colonias and districts.

4.3 Dimension and variables

Since INEGI did not include the income variable in the census of 2010, indicators of material goods and services in substitution of the income received are used as an approximation of the economic dimension, see Table (1).

Table 1 Variables measuring the economic dimension. Elaborated based on INEGI (2010)

Indicators	**Definition**
RelVPH_TV	Log ratio of private households owning a television and private households without a television.
RelVPH_RE	Log ratio of private households owning a refrigerator and private households without a refrigerator.
RelVPH_LA	Log ratio of private households owning a washing machine and private households without a washing machine.
RelVPH_AU	Log ratio of private households owning an automobile and private households without an automobile.
RelVPH_PC	Log ratio of private households owning a computer and private households without a computer.
RelVPH_TE	Log ratio of private households owning a telephone and private households without a telephone.
RelVPH_CE	Log ratio of private households owning a cellphone and private households without a cellphone.
RelVPH_IN	Log ratio of private households with internet service and private households without internet service.

5 Results

It should be noted that the census information regarding to the 13,520 urban blocks has been scaled up and grouped in the different 395 colonias. Likewise, the colonias are differentiated by utilizing the seven districts to which they belong. The representation of both scales (colonias and districts) allow a better understanding of their behavior based on the different set of dimensions.

5.1 Principal Component Analysis

A high correlation among independent variables creates problems such as multicollinearity. Accordingly, a reduction of dimensions is performed through a Principal Component Analysis (PCA). The PCA preserves as much as possible of the original structure of the data. Furthermore, this analysis generates and defines a new set of independent dimensions by taking the data set and looking for directions explaining the greatest amount of variability, see Table (2).

Table 2 Importance of components

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	2.17	1.61	0.61	0.35	0.27	0.22	0.13	0.08
Proportion of variance	59.10	32.74	4.66	1.613	0.97	0.61	0.22	0.09
Cumulative proportion	59.10	91.84	96.50	98.11	99.08	99.69	99.91	100.00

Since PC1 and PC2 captures a cumulative proportion of the 91.84 % of the variability of the information, a biplot with these two components is performed. The biplot allows to identify the behavior of the colonias around the different indicators, it helps to explain the relationship between them, and at the same time, it identifies potential clusters. See Figure (1).

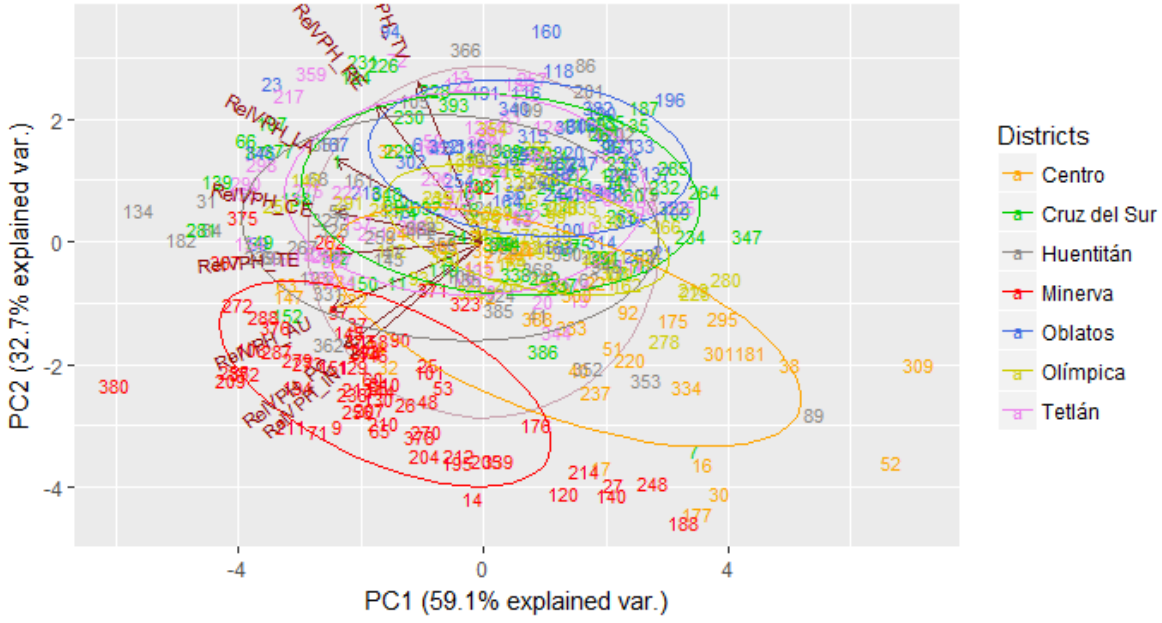


Figure 1 Principal Component biplot explaining 91.83 % of the variability of observations by colonias and urban districts.

From the biplot it is observed that the colonias belonging to the Minerva district are characterized by having a similar behavior between them and opposite from the rest. Particularly, it is observed that in the Minerva district and its colonias, the households

with the most significant proportion of cars, internet and computers are found. On the other hand, it is observed that the colonias belonging to the Oblatos district have the lowest percentage of households with cars, computers and internet access. Moreover, these three indicators have the strongest relationship with each other.

Although the rest of the colonias are little appreciated, this first approach allows observing a clear differentiation between the Minerva and Oblatos districts, a fact that highlights the west-east dichotomy in the city regarding the variables selected in the economic dimension. For this reason, in order to clearly differentiate the rest of the colonias, a hierarchical cluster analysis has been carried out.

5.2 Hierarchical cluster analysis

In this study, an agglomerative clustering is performed through hierarchical cluster analysis. Initially, each observation is assigned to its own cluster, and then the algorithm is performed iteratively, in each stage joining the two most similar clusters until there is only one cluster left. The method of Ward is used to perform the clustering of the observations. Through the dendrogram obtained from the method, it has been decided to classify the colonias into four clusters, see Figure (2).

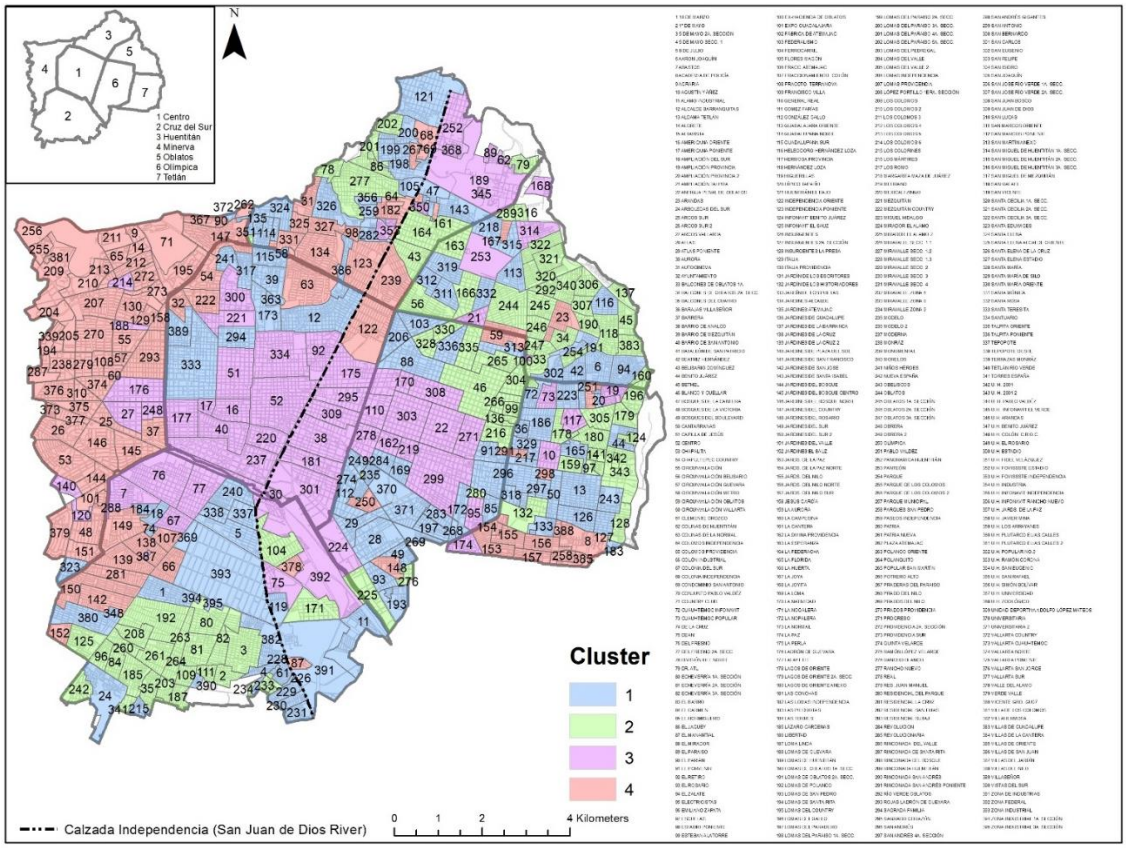


Figure 2 Colonias of Guadalajara by Cluster.

Figure (2), allows us to observe how cluster 4 is grouped mostly to the west of the Independencia Causeway, specifically in the Minerva district. On the other hand, it can be seen that within the Centro district most of the colonias are classified in cluster 3. To understand the behavior of the log ratios of the indicators in the territory, Figure (3) shows the characterization of these clusters

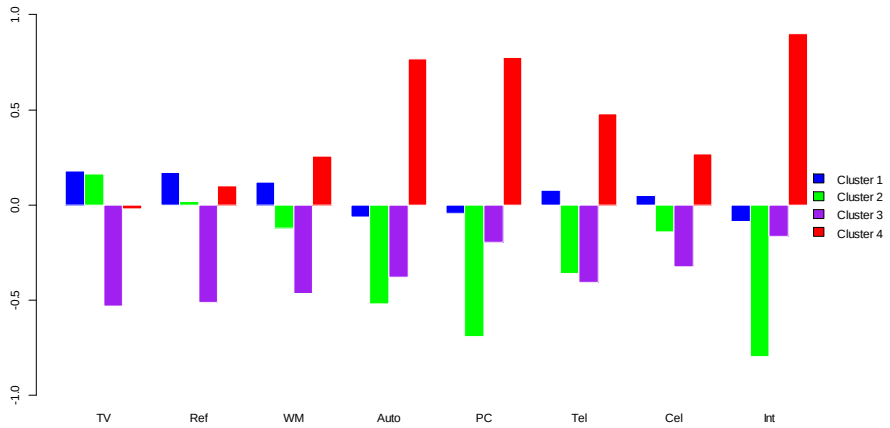


Figure 3 Cluster characterization.

From the cluster characterization, Figure (3), and based on the variables selected to measure the economic dimension, it can be seen that colonias in cluster 2 and 3 are the most disadvantaged. These clusters are characterized by dwellings that mostly lack material goods (e.g., Television, Refrigerator, Washing Machine, Car, Computer, Telephone, Cell Phone and Internet). Especially cluster 3, which lacks mostly of computer and internet access. Concerning cluster 1, its proportion of dwellings with some material goods is almost similar to the ones lacking these goods. Finally, cluster 4 stands out from the rest. The relation of households owning a car is by far the highest. Likewise, this cluster has the highest relation of households owning different goods with the exception of dwellings with television and refrigerator.

6 Discussion and conclusion

A descriptive multivariate statistical analysis of different indicators based in households goods and based on Aitchison log ratio approach (1986) with simple two-part compositions showed a first picture of the existing socio-economic segregation pattern in the city of Guadalajara. The results obtained are consistent with the patterns of the phenomenon of segregation that existed and transcended the Colonial period in the city of Guadalajara. The latter might mean that the old natural barrier of the San Juan de Dios River that separated the Spanish city from the city of Indians, and now Independencia Causeway, could have remained and become an imaginary barrier that continues dividing the rich and the poor.

The above is clearly seen in the polarization between the Minerva district and the rest of the districts. Moreover, recognizing segregation as a complex and multidimensional

phenomenon, different indicators and dimensions should be incorporated into the study.

References

Aceves, Jorge, Renée de la Torre, y Patricia Safa. «Fragmentos urbanos de una misma ciudad: Guadalajara.» *Espiral, Estudios sobre estado y Sociedad*, 2004: 277-320.

Aitchison, J. *The Statistical Analysis of Compositional Data. Monographs on Satistics and Applied Probability.* London: Chapman & Hall Ltd,, 1986.

Alvizo, Cristina. «La colonia Obrera y la segregación urbana en Guadalajara.» *Historia 2.0. Conocimiento histórico en clave digital* , 2013: 9-26.

Barajas, Luis Felipe Cabrales, y Mercedes Arabela Chong Muñoz. «Divide y venderas: Promoción inmobiliaria del barrio de artesantos de Guadalajara,1898-1908.» *Scripta Nova*, 2006: 1-17.

Brun, Jacques. «Essai critique sur la notion de ségrégation et sur son usage en géographie urbaine. .» *L´Harmattan*, 1994: 21-58.

Cabrales, Luis Felipe, y Elia Canosa. «Segregación residencial y fragmentación urbana: Los fraccionamientos cerrados en Guadalajara.» *Espiral. Estudios sobre Estado y Sociedad*, 2001: 223-253.

Consejo Nacional de Población. «Índices de Marginación.» *Secretaría de Gbernación.* 2010. http://www.conapo.gob.mx/es/CONAPO/Indices_de_Marginacion_Publicaciones.

Cowgill, D.O., y M.S. Cowgill. «An Index of Segregation Based on Block Statistics.» *American Sociological Review, vol.20, núm.2*, 1951: 825-831.

Dawkins, C.J., M. Reibel, y D.W. Wong. «Introduction-further innovations in segregation and neighborhood change research.» *Urban Geography*, 2007: 513-515.

Doñán, Juan José. *Oblatos-Colonias; Andanzas Tapatías.* Guadalajara, Jalisco: Arlequín, 2013.

Donat, Carlos. «La segregación urbana: Marco Teórico Conceptual y Estado de la Cuestión.» En *Barrios y Crisis. Crisis Económica, segregación urbana e innovación social en Cataluña*, de Ismael Blanco y Oriol Nel lo, 27-50. Valencia, Spain: Tirant Humanidades, 2018.

Donat, Carlos. «La segregación urbana: Marco Teórico Conceptual y Estado de la Cuestión.» En *Barrios y Crisis. Crisis Económica, segregación urbana e innovación social en Cataluña*, de Ismael Blanco y Oriol Nel lo, 27-50. Valencia: Tirant Humanidades, 2018.

Duncan, O.B., y B Duncan. «A Methodological Analyisis of Segregation Indexes.» *American Sociological Review, vol. 20, núm. 2*, 1955: 210-217.

Gobierno de Guadalajara. *Mapa Guadalajara.* 4 de 4 de 2019. https://mapa.guadalajara.gob.mx/geomap#0.

INEGI. *Censos y Conteos de Población y Vivienda.Insituto Nacional de Estadística y Geografía.* 31 de 12 de 2010. https://www.inegi.org.mx/programas/ccpv/2010/default.html.

Instituto Nacional de Estadística y Geografía. *Biblioteca Digital de Mapas.* 31 de 12 de

2010. https://www.inegi.org.mx/app/buscador/default.html?q=guadalajara#tabMCcollapse-Indicadores.

Jahn, Julius, Calvin F. Schmid, y Clarence Schrag. «The Measurement of Ecological Segregation.» *American Sociological Review, vol.12, núm.3*, 1947: 293-303.

Jiménez, Eduard, y Carles Donat. «El Estudio de la segregación urbana: estrategia metdológica.» En *Barrios y Crisis; Crisis económica, segregación urbana e innovación social en Cataluña*, de Ismael Blanco y Nel lo Oriol, 51-70. Valencia, Spain.: Tirant Humanidades, 2018.

Johnston, Ron, Michael Poulsen, y James Forrest. «Segregation matters, measurement matters.» En *Social-Spatial Segregation. Concepts, Processes and Outcomes*, de Christopher D. Lloyd, Ian Shuttleworth y David W. Wong, 13-44. Bristol. UK: Bristol Universty Press, 2015.

Kempen, Ronald Van. «The Academic Formulations: Explanations for the Partitioned City.» En *Of States and Cities; The Partitioning of Urban Space*, de Peter Marcuse y Ronald van Kempen, 35-56. New York: Oxford University Press, 2002.

Kertzer, D.I., y D. Arel. *Census and identity: the politics of race, ethnicity, and language in national censuses.* Cambridge: Cambridge University Press, 2002.

Lieberson, Stanley. «An Asymmetrical Approach to Segregation.» En *Ethnic segregation in Cities*, de Ceri Peach, Vaughn Robinson y Susan Smith, 61-82. Croom Helm, 1981.

Lloyd, Christopher D., Gemma Catney, y Ian G.Shuttleworth. «Measuring neighbourhood segregation using spatial interaction data.» En *Social-Spatial Segregation. Concepts, Processes and Outcomes*, de Christopher D. Lloyd, Ian G.Shuttleworth y David W. Wong., 65-90. Bristol,UK: Bristol University Press, 2015.

Lloyd, Christopher D., Ian Shuttleworth, y David W. Wong. *Social-Spatial Segregation. Concepts, Processes and outcomes.* Bristol, UK: Policy Press, 2015.

López Moreno, Eduardo. *La cuadrícula en el desarrollo de la ciudad hispanoamericana. Guadalajara, México.* Guadalajara, Jalisco, México.: Universidad de Guadalajara. Instituto Tecnológico y de Estudios Superiores de Occidente, 2001.

Marrufo, Rafael Mauricio, y Sonia. Bass Zavala. «Segregación socioespacial y servicios de salud en Ciudad Juárez.» En *Segregación Urbana y Espacios de Exclusión; Ejemplos de México y América Latina*, de Adrián Guillermo Aguilar y Irma Escamilla H., 139-162. México: Universidad Nacional Autónoma de México. Miguel Ángel Porrúa, 2015.

Massey, Douglas S., y Nancy A. Denton. «The Dimension of Residential segregation.» *Social Forces, Vol.67, No.2* , 1988: 281-315.

Mateos, Pablo. «The international comparability of ethnicity and collective identity: implications for segregation studies.» En *Social-Spatial Segregation. Concepts, Processes and Outcomes*, de Christopher D. Lloyd, Ian G.Shuttleworth y David W. Wong., 163-196. Bristol,UK: Bristol University Press, 2015.

Moreno Pérez, Orlando. «Insustentabilidad de la vida, segregación social y pobreza urbana: efectos de las políticas de vivienda en la era del ultraliberalismo.» En

Segregación Urbana y Espacios de Exclusión; Ejemplos de México y América Latina, de Adrián Guillermo Aguilar y Irma Escamilla H., 307-328. México: Universidad Nacional Autónoma de México. Miguel Ángel Porrúa, 2015.

Pawlowsky-Glahn, V, y J.J. Egozcue. «Compositional data and their analysis: an introduction.» En *Compositional Data Anlaysis in the Geosciences*, de A Buccianti, G Mateu-Figueras y V. Pawlosky-Glahn, 1-10. London: The Geological Society, 2006.

Peach, C. «The meaning of segregation.» *Planning Practice and Research, vol. 11, no 2*, 1996: 137-150.

Rivera, Elizabeth. «Introspecciones sobre el proceso, producción, construcción y reconstrucción de nuevos espacios de identificación urbana, a través de la memoria, vivencias y experiencias de sus habitantes. El caso del Barrio de San Juan de Dios en la ciudad de Guadalajara.» *Roma*, 2013: 1133-1142.

—. «Transformación socio-espacial y dinámicas del uso del suelo en Guadalajara, México: Análisis la producción del espacio urbano-metropolitano y sus posibles escenarios.» Centre de Política de Sòl i Valoracions Universidade Federal do Rio de Janeiro, 2012. 1-19.

Romero Mares, Patricia Isabel, y Josefina Hernández Lozano. «Propuesta de metodología para analizar el nivel de segregación residencial en la Zona Metropolitana de México.» En *Segregación Urbana y Espacios de Exclusión; Ejemplos de México y América Latina*, de Adrián Guillermo Aguilar y Irma Escamilla H., 223-239. México: Universidad Nacional Autónoma de México. Miguel Ángel Porrúa, 2015.

Schteingart, Martha. «La división social del espacio en ciudades mexicanas: un balance explicativo desde una perspectiva latinoamericana.» En *Segregación urbana y espacios de exclusión. Ejemplos de México y América Latina.* , de Adrián Guillermo Aguilar y Irma Escamilla H., 47-72. Ciudad de México: Miguel Ángel Porrúa, 2015.

Subirats, Joan. «Pobreza y exclución social. Un análisis de la realidad española y europea.» *Fundación "La Caixa"*, 2004: 1-190.

Vázquez, Daniel. *Guadalajara: Ensayos de interpretación.* Guadalajara: El Colegio de Jalisco, 1989.

Wong, David S. «Using general spatial pattern statistic to evaluate spatial segregation.» En *Social-Spatial Segregation. Concepts, Processes and Outcomes*, de Christopher D. Lloyd, Ian Shuttleworth y David W. Wong, 45-64. Bristol, UK: Bristol University Press, 2015.

Partial Correlations in Compositional Data Analysis

Ionas Erb[1]
[1]Centre for Genomic Regulation (CRG),
The Barcelona Institute of Science and Technology,
Barcelona, Spain; *ionas.erb@crg.eu*

Summary

Partial correlations quantify linear association between two variables adjusting for the influence of the remaining variables. They form the backbone for graphical models and are readily obtained from the inverse of the covariance matrix. For compositional data, the covariance structure is specified from log ratios of variables, so unless we try to "open" the data via a normalization, this implies changes in the definition and interpretation of partial correlations. In the present work, we elucidate how results derived by Aitchison (1986) lead to a natural definition of partial correlation that has a number of advantages over current measures of association. For this, we show that the residuals of log-ratios between a variable with a reference, when adjusting for all remaining variables including the reference, are reference-independent. Since the reference itself can be controlled for, correlations between residuals are defined for the variables directly without the necessity to recur to ratios except when specifying which variables are partialled out. Thus, perhaps surprisingly, partial correlations do not have the problems commonly found with measures of pairwise association on compositional data. They are well-defined between two variables, are properly scaled, and allow for negative association. By design, they are subcompositionally incoherent, but they share this property with conventional partial correlations (where results change when adjusting for the influence of fewer variables). We discuss the equivalence with normalization-based approaches whenever the normalizing variables are controlled for. We also discuss the partial variances and correlations we obtain from a previously studied data set of Roman glass cups.

Key words: Compositional covariance structure, inverse log-ratio covariance, residual log-ratio variance, compositional pairwise association, partial proportionality.

1 Introduction

1.1 Background and outline

Since the publication of Aitchison's book on compositional analysis (Aitchison, 1986), awareness has increased about the fact that correlations between variables are problematic when these variables are parts of compositional data. To remedy the problems, a number of alternatives to correlation have been suggested. The most prominent is log-ratio variance (Aitchison, 1986), but there are also bounded (Aitchison, 2003) or scaled (Lovell et al., 2015) versions of it. All of these measures have their own problems: one would like to judge the value of log-ratio variance according to the intrinsic variability of its constituents, not only bound it between zero and one. Scalings however can lead to spurious results that become apparent when the underlying absolute data are known (Erb and Notredame, 2016). Apart from this, negative associations

ISBN 978-84-947240-2-2

remain beyond the scope of all these measures of proportionality. Interestingly, the correlations between log-ratios that use the geometric mean as a reference, i.e., in clr-transformed data, have found little acceptance although they are perhaps the solution coming closest to a genuine correlation for compositional data. Its main drawback is subcompositional incoherence. However, the insistence on subcompositional coherence, and perhaps a certain reluctance to the use of correlation in the community, seem to have blocked the way towards embracing *partial* correlations as a valid way of analyzing pairwise association between variables summing to a constant. Clearly, here the coherence of results obtained on subsets of variables is neither required nor possible: all partial correlations change when removing variables we control for. To the best of our knowledge, this fact has not been exploited in compositional analysis, and work on partial correlation is lacking. This is especially surprising as the necessary mathematical results were derived by Aitchison in the 1980s already.

Indeed, when considering pairwise association between variables while adjusting for dependence on all other variables including the reference, the correlation that remains seems the best we can expect from a linear pairwise association measure of compositional parts. In this paper, we use some results from Aitchison (1986) to derive that correlating log-ratios when their reference variables are controlled for makes the correlations reference-independent in the sense that any reference from the variables we control for can be used interchangeably without altering the result. For a simple example, take the case of four compositional random variables $X_1, \ldots, X_4$. Let us denote by

$$r_{1,2|3,4}(X_1, \ldots, X_4) = \operatorname{corr}\left(\log\frac{X_1}{X_4}, \log\frac{X_2}{X_4}\,\middle|\,\log\frac{X_3}{X_4}\right) \tag{1}$$

the partial correlation between the logarithms of the first two variables referenced by X_4 adjusting for the influence of the log-ratio X_3 with the reference. The notation on the left-hand side seems ambiguous but it turns out that

$$\begin{aligned} r_{1,2|3,4}(X_1, \ldots, X_4) &= \operatorname{corr}\left(\log\frac{X_1}{X_3}, \log\frac{X_2}{X_3}\,\middle|\,\log\frac{X_3}{X_4}\right) \\ &= \operatorname{corr}\left(\log\frac{X_1}{\sqrt{X_3X_4}}, \log\frac{X_2}{\sqrt{X_3X_4}}\,\middle|\,\log\frac{X_3}{X_4}\right) = \operatorname{corr}\left(\log\frac{X_1}{X_3}, \log\frac{X_2}{X_4}\,\middle|\,\log\frac{X_3}{X_4}\right). \end{aligned} \tag{2}$$

Although partial correlations of the form (1) are standard (in the sense that the theory of partial correlations applies directly to log ratios), it takes some results from the statistics of compositional data analysis to show that there is no dependence on the reference variable as long as it is included in the variables we control for. Thus partial correlations of log ratios are simpler than they appear at first sight, with some interesting implications. First, the number of correlations to interrogate is restricted to pairs and is thus drastically reduced with respect to the case where references have to be taken into account. Second, attempts to "open" the data, i.e. to analyze them in absolute terms after normalization with an unchanged reference appear of little use if for most cases the same results can be derived without normalization. Third, the covariance matrix of the geometric-mean-referenced parts (i.e., of the clr-transformed data) can be used via its pseudoinverse to obtain all necessary results.

1.2 Prerequisites

1.2.1 Data matrices as instances of compositional random vectors

We start with a hypothetical, real-valued but positive $N \times D$ data matrix with elements a_{ij}, where samples are indexed in the rows and variables in the columns. The N rows of this matrix can be considered instances of a D-dimensional random vector $A = (A_1, \ldots, A_D)$. In order to avoid problems with zeros when dealing

with log ratios, we also assume $A_j > 0$ throughout. Let us now consider the closure of these instances, i.e. the data matrix (x_{ij}) resulting when we divide each element a_{ij} by its row sum. The corresponding random vector $X = A/\sum(A_j)$ we call a compositional random vector. Compositional analysis concerns data where the total sum over a sample has no relevance to the analyst. This sum can vary between samples. The closure operation is just a convenient way of incorporating the fact that the total sum over the constituent random variables is arbitrary. The "absolute" data a_{ij} are considered unavailable here and were only mentioned to clarify that they can underlie x_{ij}. Data other than a_{ij} can of course lead to the same x_{ij}. Put another way, we use the random vector X as our representative for the equivalence class of random vectors with the same relative relationships between their random variables.

Let us now introduce yet another kind of random vector, Z, which we call the log-ratio transformed random vector. It is defined by $Z = \log(X/g(X))$, where g denotes the geometric mean over the variables in the random vector. We can restrict the variables to an index set $\mathcal{A}$ and write this projection as $X_{\mathcal{A}}$. The resulting denominator $g(X_{\mathcal{A}})$ is called the reference. In the case of $\mathcal{A}$ containing all indices of the composition under consideration, the transformation is called the centered log-ratio transformation (clr), and the resulting random vectors we denote by Z, or $Z_{\mathcal{A}}$, whatever subset is considered. Throughout the manuscript, we let the set indexing Z also indicate the variables over which the geometric mean is taken (unless a single variable is indicated, like for Z_j). In the case of the reference containing a single variable, usually X_D, the log-ratio transformation is called additive (alr). Note that the use of the index D is a matter of convenience, and a variable of interest can just be moved to this last position. In general, results will depend on this choice. We will denote these alr-transformed random vectors by Y, and their restrictions to a subset $\mathcal{A}$ as $Y_{\mathcal{A}}$.

The two types of log-ratio transformed random vectors can be transformed into each other using a matrix designed for this purpose (Aitchison, 1986). Let us denote by $\mathbf{I}_d$ the d-dimensional identity matrix and by $\mathbf{j}_d$ the d-dimensional vector with entries 1. Let us define the matrix $\mathbf{F} = [\mathbf{I}_{D-1}, -\mathbf{j}_{D-1}]$, i.e. $\mathbf{F}$ is the $(D-1)\times D$ matrix resulting from writing $\mathbf{I}_{D-1}$ and $-\mathbf{j}_{D-1}$ side by side. It is then easy to verify that we have

$$Y = \mathbf{F}Z. \tag{3}$$

1.2.2 Compositional covariance specifications and inverse variance

Aitchison (1986) introduced three different matrices, each (equivalently) specifying the full covariance structure of a compositional data set. The first matrix has the log-ratio variances $\mathrm{var}(\log(X_i/X_j))$ as elements and will not concern us further. The other two specifications are covariance matrices of the log-ratio transformed random vectors. With the notation from the previous section, the $(D-1)\times(D-1)$ matrix of alr-transformed random vectors is $\mathbf{\Sigma} = (\sigma_{ij}) = (\mathrm{cov}(Y_i, Y_j)) = \mathrm{var}(Y)$. Finally, the third matrix is the covariance of clr-transformed random vectors $\mathbf{\Gamma} = (\gamma_{ij}) = (\mathrm{cov}(Z_i, Z_j)) = \mathrm{var}(Z)$. The change in log-ratio transformation results in a singular $D \times D$ matrix, with the singularity resulting from the constraint $\sum_j Z_j = 0$. Similarly to (3), we can also transform the covariance matrices into each other:

$$\mathbf{\Sigma} = \mathbf{F}\mathbf{\Gamma}\mathbf{F}^T. \tag{4}$$

With partial correlations in mind, an important result derived by Aitchison (1986) concerns the relationship between the pseudoinverse $\mathbf{\Gamma}^-$ of $\mathbf{\Gamma}$ and the inverse variance $\mathbf{\Sigma}^{-1}$. We have

$$\mathbf{\Gamma}^- = \mathbf{F}^T\mathbf{\Sigma}^{-1}\mathbf{F} \tag{5}$$

(see Property 5.6 (a) in Aitchison, 1986). In this article, we will essentially elucidate the implications of this relationship.

1.2.3 Partial correlation: The standard setting

Due to its unconstrained nature, the covariance matrix $\mathbf{\Sigma}$ can be used to obtain partial correlations (between log ratios having X_D as a reference) in the standard way. Let us quickly review this procedure (see, e.g., Whittaker, 1990). In the following, to achieve more economical expressions, we assume that all log-ratio transformed random variables are centered, i.e. their averages are zero (in case they are not, we just have to subtract their average from them). Let the random vector $Y_\mathcal{C}$ be composed of the set of random variables having indices in $\mathcal{C}$, i.e. $(Y_i)_{i\in\mathcal{C}}$. The linear least squares predictor (LLSP) of Y_j given $Y_\mathcal{C}$ is then defined by

$$\hat{Y}_j(Y_\mathcal{C}) = \mathrm{cov}(Y_j, Y_\mathcal{C})\mathrm{var}(Y_\mathcal{C})^{-1}Y_\mathcal{C}. \tag{6}$$

Here, $\mathrm{cov}(Y_j, Y_\mathcal{C})$ is the row vector containing the covariances of Y_j with the scalar random variables in $Y_\mathcal{C}$, and $\mathrm{var}(Y_\mathcal{C})^{-1}$ is the inverse covariance matrix of these random variables. Note that, as a consequence of this definition, $\hat{Y}_j$ lies in the space spanned by $Y_\mathcal{C}$ and the residual $Y_j - \hat{Y}_j(Y_\mathcal{C})$ is orthogonal to that space.[1] The residual variance

$$\mathrm{var}(Y_j|Y_\mathcal{C}) = \mathrm{var}\left(Y_j - \hat{Y}_j(Y_\mathcal{C})\right) \tag{7}$$

is a useful summary statistic that tells us how well Y_j can be predicted from the variables $Y_\mathcal{C}$. It is also known as partial variance. More generally, the partial covariance between two scalar random variables is defined as the covariance between their residuals:

$$\mathrm{cov}(Y_i, Y_j|Y_\mathcal{C}) = \mathrm{cov}\left(Y_i - \hat{Y}_i(Y_\mathcal{C}), Y_j - \hat{Y}_j(Y_\mathcal{C})\right). \tag{8}$$

Partial correlations are obtained scaling by the corresponding residual variances in the standard way. There is, however, a shortcut to this. An important result states that partial variances and correlations adjusting for all the remaining variables can be obtained from the inverse covariance matrix. We have

$$\mathrm{var}(Y_j|Y_{\{1,\ldots,D-1\}\setminus j}) = 1/\sigma^{(-1)}_{jj}, \tag{9}$$

$$\mathrm{corr}(Y_i, Y_j|Y_{\{1,\ldots,D-1\}\setminus\{i,j\}}) = \frac{-\sigma^{(-1)}_{ij}}{\sqrt{\sigma^{(-1)}_{ii}\sigma^{(-1)}_{jj}}}, \tag{10}$$

where $\sigma^{(-1)}_{ij}$ denote the elements of $\mathbf{\Sigma}^{-1}$.

2 Partial correlation on compositional data

2.1 The residual is independent of the choice of alr transformation

While $\mathbf{\Sigma}$ is a standard covariance matrix, in a compositional context its usefulness may not be apparent as it depends on the reference chosen. Rather surprisingly, it turns out that the inverse $\mathbf{\Sigma}^{-1}$ is essentially independent of the choice of reference. This is implied by (5), but although the equation may seem unsurprising given (4), the fact that a matrix containing all the parts can be obtained from one where a part is sacrificed as reference appears somewhat mysterious. This can be understood better when looking from another angle. Remember that inverse covariance matrices have diagonal elements related to the variance of the residuals, see (9). The fact that for $\mathbf{\Sigma}^{-1}$ the choice of reference has little importance can be made sense of if the residuals themselves are independent of this choice.

[1] More precisely, the predictor and the residual generate N-sample vectors that lie in/are orthogonal to the space spanned by the N-sample vectors generated from the random variables in $Y_\mathcal{C}$.

 ISBN 978-84-947240-2-2

Remember that Y_j is an alr-transformed random variable with reference X_D and that we have $\mathbf{\Sigma} = \mathrm{var}(Y)$ with Y having $D-1$ components. For removing Y_j form Y, let us define the index set $\mathcal{C}_j = \{1, \ldots, D-1\} \backslash j$. We now show that the residual

$$Y_j - \hat{Y}_j \left(Y_{\mathcal{C}_j} \right) \tag{11}$$

remains the same irrespective of which reference is chosen from the set $\{X_k | k = 1, \ldots, D, k \neq j\}$ for constructing its constituent random variables $\log X_i/X_k$, $i \neq j, k$. For this, we first have to clarify that the space generated from the explanatory log ratios (the random vector $Y_{\mathcal{C}_j}$ we are adjusting for) remains unchanged when choosing another variable for reference. Intuitively, a change of reference in the variables we adjust for corresponds to an overall subtraction of one of these variables (the one that has the new reference in the numerator) and neither reduces nor moves us outside the space generated by $Y_{\mathcal{C}_j}$. As prediction is equivalent with linear projection into this space, the LLSP will remain the same. A formal proof that permuting the variables in $Y_{\mathcal{C}_j}$ does not change the LLSP is presented in the Appendix.[2] Now let us come back to the residual. A change from reference X_D to reference X_k is achieved by subtracting Y_k from Y_j ($k \neq j, D$):

$$Y_j - Y_k - (\widehat{Y_j - Y_k}) = Y_j - \hat{Y}_j - (Y_k - \hat{Y}_k), \tag{12}$$

where the equality comes form the linearity of the predictor. With the definition of the LLSP (6) and denoting the elements of $\mathrm{var}(Y_{\mathcal{C}_j})^{-1}$ by $c_{il}^{(-1)}$ we have

$$\hat{Y}_k \left(Y_{\mathcal{C}_j} \right) = \mathrm{cov}(Y_k, Y_{\mathcal{C}_j}) \mathrm{var}(Y_{\mathcal{C}_j})^{-1} Y_{\mathcal{C}_j} = \left(\sum_{i \in \mathcal{C}_j} \sigma_{ki} c_{il}^{(-1)} \right)_{l \in \mathcal{C}_j}^{T} Y_{\mathcal{C}_j} = (\delta_{kl})_{l \in \mathcal{C}_j}^{T} Y_{\mathcal{C}_j} = Y_k. \tag{13}$$

Here, δ denotes the Kronecker delta. It appears because the elements of $\mathbf{\Sigma}$ and of $\mathrm{var}(Y_{\mathcal{C}_j})$ coincide for indices in $\mathcal{C}_j$. Comparing with the right-hand side of (12), we see that the term in brackets vanishes, thus establishing the independence of the residual from a reference X_k, $k \neq j$. Figure 1 shows an intuitive rendering of the argument we provided. For the argument it is crucial that the reference is coming from the constituent variables of the ratios we adjust for.

2.2 Connection with clr transformation

We have seen that our residuals do not depend on the particular alr transformation chosen (as long as the reference occurs as a variable in the ratios we are adjusting for). What about the reference coming from a clr transformation? Note that for an equivalent formulation, we need to consider the clr on the subspace of variables we are adjusting for. A similar argument as in the previous section (see Appendix) shows that an equivalent expression of our residual (11) that is symmetric with respect to the reference variables is

$$Z_j - \hat{Z}_j \left(Z_{\mathcal{D}_j} \right), \tag{14}$$

where the geometric mean reference of Z_j goes over the indices in $\mathcal{D}_j = \{1, \ldots, D\} \backslash j$. Now from (5) we know that the diagonal elements of the pseudoinverse of $\mathbf{\Gamma}$ are

$$\gamma_{jj}^{(-)} = \sigma_{jj}^{(-1)}, j = 1, \ldots, D-1. \tag{15}$$

Due to the symmetry of $\mathbf{\Gamma}$ with respect to reference, this also has to hold after permuting labels when constructing $\mathbf{\Sigma}$ (see Appendix for a proof). Thus, with (9), we conclude that for all indices $j = 1, \ldots, D$,

[2] Note that this does not imply that the LLSP is independent of the choice of reference, i.e., in general we have $\log(\widehat{X_j/X_D}) \neq \log(\widehat{X_j/X_k})$.

 ISBN 978-84-947240-2-2

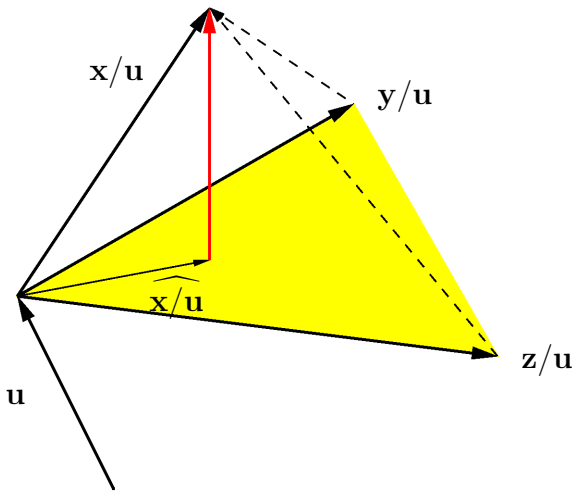


Figure 1: The residual of $\mathbf{x}$ is independent of its reference. Shown is the $D = 4$ case with (logged) vectors $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ and $\mathbf{u}$, the latter serving as a reference. To simplify, logs are omitted in the vector annotations. The linear least squares predictor of $\log(\mathbf{x}/\mathbf{u})$ with respect to $(\log(\mathbf{y}/\mathbf{u}), \log(\mathbf{z}/\mathbf{u}))$ is shown as the thin arrow lying in the yellow plane. The residual vector is shown in red. It can be seen that replacing $\log(\mathbf{x}/\mathbf{u})$ by $\log(\mathbf{x}/\mathbf{y})$ or $\log(\mathbf{x}/\mathbf{z})$ (dashed lines) results in the same residual.

the residual variance can be calculated from

$$\mathrm{var}\left(Z_j \,\middle|\, Z_{\mathcal{D}_j}\right) = 1/\gamma_{jj}^{(-)}. \tag{16}$$

Similarly, the off-diagonal elements of $\mathbf{\Gamma}^-$ coincide with the ones of $\mathbf{\Sigma}^{-1}$ when they have indices $i, j < D$. Again invoking invariance with respect to label permutations, we conclude that for all $i, j = 1, \ldots, D$, partial correlations can be calculated from

$$\mathrm{corr}\left(Z_i, Z_j \,\middle|\, Z_{\mathcal{D}_{ij}}\right) = \frac{-\gamma_{ij}^{(-)}}{\sqrt{\gamma_{ii}^{(-)}\gamma_{jj}^{(-)}}}, \tag{17}$$

where we used the notation $\mathcal{D}_{ij} = \{1, \ldots, D\}\backslash\{i, j\}$. Note that here the Z_i, Z_j are different from the ones in (16) because the geometric mean reference is taken over $X_{\mathcal{D}_{ij}}$, not $X_{\mathcal{D}_j}$.

2.3 A general expression, correlation matrices, R^2

To summarize, we have shown that we can use as a reference either variable occurring in the explanatory ratios or the geometric mean taken over them. For compositional data it is thus unambiguous to define partial variances and correlations directly for the parts X_i of the composition via

$$\sigma^2_{j|\mathcal{D}_j}(X) = 1/\gamma_{jj}^{(-)}, \tag{18}$$

$$r_{ij|\mathcal{D}_{ij}}(X) = \frac{-\gamma_{ij}^{(-)}}{\sqrt{\gamma_{ii}^{(-)}\gamma_{jj}^{(-)}}}, \tag{19}$$

where we mean the variances and correlations evaluated on residuals of the form (11) or (14) and with any reference coming from the log ratios we are adjusting for.

Correlation matrices are easier to interpret than covariance matrices, and the inverses of correlation matrices are often used in practice. Although $\mathbf{\Sigma}^{-1}$ and $\mathbf{\Gamma}^{-}$ coincide where they can, their correlation counterparts do not. Indeed, when scaling $\mathbf{\Sigma}$ and $\mathbf{\Gamma}$ to have unit entries in the diagonal,[3] after inverting the scaled version of $\mathbf{\Sigma}$ and pseudo-inverting the scaled version of $\mathbf{\Gamma}$, we have to re-scale both to unit diagonal to make them coincide again. Their multiple correlation coefficients R^2, however, remain different. As they quantify the amount of variance of Y_j or Z_j explained by their respective predictors (which depend on the reference), they have to be reference-dependent:

$$\mathrm{var}(\hat{Y}_j)/\mathrm{var}(Y_j) = 1 - 1/\left(\sigma_{jj}\sigma_{jj}^{(-1)}\right), \tag{20}$$

$$\mathrm{var}(\hat{Z}_j)/\mathrm{var}(Z_j) = 1 - 1/\left(\gamma_{jj}\gamma_{jj}^{(-)}\right). \tag{21}$$

Although Y_j and Z_j have the same partial variance $\sigma^2_{j|\mathcal{D}_j}(X) = 1/\gamma_{jj}^{(-)} = 1/\sigma_{jj}^{(-1)}$, their zero-order variances are different. When an R^2 needs to be reported, (21) would be a likely candidate or otherwise one would have to have a good reason to choose a more specific reference X_D.

2.4 Comparison with normalization-based partial correlations

An approach that is sometimes pursued (e.g., in genomics) consists in trying to recover the underlying absolute data a_{ij} mentioned in section 1.2.1. This is achieved by means of a specific normalization, i.e., by multiplication of the rows of (x_{ij}) with factors s_i that are proportional to $\sum_j a_{ij}$. These factors are unknown in principle (they got lost by the closure operation). They can sometimes be inferred from assumptions concerning the knowledge of (on average) unchanged a_{ij} across rows (see the supplement to Quinn et al., 2018). If the assumptions are fulfilled, the resulting data matrix is no longer relative and there is no need for analyzing ratios. Note however that we can write the normalized data as a special case of log-ratio transformed data $Z = \log(X/g(X_{\mathcal{U}}))$, where the set $\mathcal{U}$ contains only indices of variables for which we have $g(A_{\mathcal{U}})$ =const. (We have just transformed to logarithms of the normalized data, a common procedure in genomics.) When comparing with the partial correlations for log-ratio transformed data with more general references (19), we see that they coincide (and can be obtained assumption-free, without normalization) as long as the log ratios we are adjusting for also contain the variables indexed by $\mathcal{U}$.

3 Application to Roman glass-cup data

We evaluate (18) and (19) on a data set of 11 oxides and elements composing the glass of 47 Roman cups excavated in Colchester (Cool and Price, 1994). We use the version of the data presented in (Baxter et al., 1990). The few previous analyses have focused on ordination. One of the conclusions of Greenacre's recent analysis (Greenacre, 2018) is that the ratio of SiO_2/CaO is the significant dimension of variation, followed by the ratios of SiO_2/Sb and Na_2O/Sb, which were found to be of lesser influence for the multivariate structure. In Table 1, beside the average weight percentages of the various oxides and elements, we show the partial variances and log-ratio variances (with respect to the geometric mean). For better readability, both of them are divided by total variance as obtained from the trace of $\mathbf{\Gamma}$. Then, R^2 with the mean reference is shown. In the last two columns we also present log-ratio variance and R^2 with respect to SiO_2. The oxides occurring in high weight percentages, with the exception of SiO_2, tend to have low variances. The relatively high variance of SiO_2 is however well explained by the other variables (highest R^2 with geometric mean reference). This

[3] Transforming a covariance into a correlation matrix is achieved by multiplying from the left and the right a suitably scaled diagonal matrix.

may explain the good properties as a reference reported by Greenacre (2018). The two variables with the highest residual variance (MnO and Sb) also contribute most variance to the data. The values of R^2 with respect to SiO_2 show that Fe_2O_3 and Al_2O_3 are best predicted using this reference. In table 2 we show the

Table 1: Single-variable quantities for Roman glass cups. All values are given in percent.

oxide, element	av. weight	res. var/tot.	var/tot. (mean)	R^2 (mean)	var/tot. (SiO_2)	R^2 (SiO_2)
SiO_2	72	0.86	7.7	89	-	-
Al_2O_3	1.9	0.37	2.5	85	2.0	90
Na_2O	18	0.75	3.6	79	1.4	70
Fe_2O_3	0.30	1.5	5.9	75	11	93
MgO	0.46	3.4	6.4	46	8.3	78
CaO	5.6	1.2	1.9	39	3.5	81
TiO_2	0.07	3.5	4.9	29	8.2	77
MnO	0.01	21	29	26	30	61
Sb	0.35	22	28	23	22	45
P_2O_5	0.05	4.3	5.3	19	8.0	70
K_2O	0.49	3.6	4.2	15	5.9	67

five strongest partial correlations as well as a plug-in estimate of their local false discovery rates obtained under permutations of the samples for each variable in the data matrix (see Appendix). Of note, permuting the residuals directly leads to a less severe test and much lower q-values, presumably because a natural dependence among the residuals gets destroyed. While Greenacre's analysis stresses the importance of SiO_2

Table 2: Partial correlations and their q-values of five pairs of variables

variable 1	variable 2	partial correlation	q-value
Fe_2O_3	Al_2O_3	0.73	0.03
Al_2O_3	SiO_2	0.72	0.03
Fe_2O_3	SiO_2	-0.66	$< 10^{-4}$
Na_2O	CaO	0.60	0.18
MgO	Fe_2O_3	0.43	0.46

and the oxides with high weight percentages for the multivariate structure, partial correlations and R^2 seem to point us to the importance of Fe_2O_3 and Al_2O_3 for the understanding of mutual dependencies. Of note is the conspicuous negative correlation between SiO_2 and Fe_2O_3, the only negative association of importance that we find in this data set. It would remain undetected with current association measures.

4 Discussion

A proposal for partial proportionality by Erb and Notredame (2016) only controls for one ratio and thus introduces an additional parameter. Already the reference can be considered a nuisance parameter and needs justification when discussing results. The natural definition of partial correlation elucidated in this paper treats the reference like just another variable that is controlled for. While we cannot quite get around discussing its choice for R^2, partial variances and partial correlations are defined reference-free and can

be evaluated as efficiently as in absolute analysis. Application to graphical models is a natural next step. For applications in genomics, e.g., the equivalence of the compositional approach with the one using a normalization of the data is interesting. The equivalence holds as long as the normalizing variables are controlled for in the sense discussed in section 2.4. In genomics it will also be necessary to modify the analysis for the many-variables setting. Here, the empirical covariance matrix is a poor estimator that can be improved by regularization, also allowing for the necessary covariance inversion.

Acknowledgements

I thank Cedric Notredame and Thom Quinn for their support and encouragement.

References

Aitchison, J (1986). *The Statistical Analysis of Compositional Data.* London: Chapman & Hall.

Aitchison, J (2003). *A concise guide to compositional data analysis. The 2nd Compositional Data Analysis Workshop; Girona, Spain.*

Baxter MJ, Cool HEM, Heyworth MP (1990). Principal component and correspondence analysis of compositional data: some similarities. *J Appl Stat 17*, pp 229–235.

Cool, HEM and Price, J (Eds.) (1994). *Colchester Archeological Report 8: Roman vessel glass from excavations in Colchester, 1971-85.* Hunstanton: Witley.

Erb, I. and Notredame, C (2016). How should we measure proportionality on relative gene expression data? *Theory in Biosciences 135*(1-2), pp 21–36.

Greenacre, M (2018). Variable Selection in Compositional Data Analysis Using Pairwise Logratios. *Mathematical Geosciences* (Online First), pp 1–34.

Hastie, T, Tibshirani, R and Friedman, J (2001) *The elements of statistical learning.* New York: Springer.

Lovell, D, Pawlowsky-Glahn, V, Egozcue, J, Marguerat, S and Bähler, J (2015). Proportionality: a valid alternative to correlation for relative data. *PLoS Computational Biology*, *11*(3).

Quinn, TP, Erb, I, Richardson, MF, Crowley, TM (2018). *Understanding sequencing data as compositions: an outlook and review*, *34*(16), pp 2870–2878.

Whittaker, J. (1990). *Graphical models in applied multivariate statistics.* Chichester: Wiley.

Appendix

Independence of LLSP from permutation of parts in the explanatory ratios

Let $\mathbf{P}$ be a $D \times D$ permutation matrix (obtained from permuting the rows of the identity matrix). When applying $\mathbf{P}$ to X, it changes the order of the random variables in X. Together with our definition of $\mathbf{F}$ from section 1.2.2, we define

$$\mathbf{Q}_P = \mathbf{F}\mathbf{P}\mathbf{F}^T\mathbf{H}^{-1}, \tag{22}$$

where $\mathbf{H}$ is defined as the $(D-1)\times(D-1)$ matrix resulting from adding a matrix of units to the identity matrix.[4] Also, it can be obtained from $\mathbf{H}=\mathbf{F}\mathbf{F}^T$. Note that, when applying $\mathbf{Q}_P$ to Y, it yields a permutation of parts in the constituent variables, e.g., for a suitable $\mathbf{P}$, $Y=(\log(X_1/X_3),\log(X_2/X_3))$ becomes $(\log(X_1/X_2),\log(X_3/X_2))=Y^P$. More formally, Property 5.2 (b) in (Aitchison, 1986) states that

$$Y^P = \mathbf{Q}_P Y, \tag{23}$$
$$\mathbf{\Sigma}_P = \mathbf{Q}_P \mathbf{\Sigma} \mathbf{Q}_P^T. \tag{24}$$

Here, $\mathbf{\Sigma}$ and $\mathbf{\Sigma}_P$ denote the covariance matrices of Y and Y_P, respectively. Let us now use versions of these matrices to contain only variables with indices from $\mathcal{C}_j=\{1,\dots,D-1\}\backslash j$. For ease of notation, we will not change the notation for $\mathbf{Q}_P$ for the lower-order version, but to use the same notation as in the main text, let us use the shorthand $\mathbf{C}=\mathrm{var}(Y_{\mathcal{C}_j})$ instead of $\mathbf{\Sigma}$. (Remember that in $Y_{\mathcal{C}_j}$, X_D occurs in the denominator of the ratios.) We can now show that $\hat{Y}_j(Y^P_{\mathcal{C}_j})=\hat{Y}_j(Y_{\mathcal{C}_j})$:

$$\begin{aligned}\hat{Y}_j(Y^P_{\mathcal{C}_j}) &= \mathrm{cov}\left(Y_j, Y^P_{\mathcal{C}_j}\right)\mathrm{var}(Y^P_{\mathcal{C}_j})^{-1}Y^P_{\mathcal{C}_j} = \mathrm{cov}\left(Y_j,\mathbf{Q}_P Y_{\mathcal{C}_j}\right)\mathbf{C}_P^{-1}\mathbf{Q}_P Y_{\mathcal{C}_j}\\ &= \mathrm{cov}\left(Y_j, Y_{\mathcal{C}_j}\right)\mathbf{Q}_P^T\mathbf{C}_P^{-1}\mathbf{Q}_P Y_{\mathcal{C}_j} = \mathrm{cov}\left(Y_j, Y_{\mathcal{C}_j}\right)\mathbf{Q}_P^T\left(\mathbf{Q}_P\mathbf{C}\mathbf{Q}_P^T\right)^{-1}\mathbf{Q}_P Y_{\mathcal{C}_j}\\ &= \mathrm{cov}\left(Y_j, Y_{\mathcal{C}_j}\right)\mathbf{Q}_P^T\left(\mathbf{Q}_P^T\right)^{-1}\mathbf{C}^{-1}\mathbf{Q}_P^{-1}\mathbf{Q}_P Y_{\mathcal{C}_j} = \mathrm{cov}\left(Y_j, Y_{\mathcal{C}_j}\right)\mathbf{C}^{-1}Y_{\mathcal{C}_j} = \hat{Y}_j(Y_{\mathcal{C}_j}).\end{aligned} \tag{25}$$

Here, the first line uses the definition of the LLSP (6), the identity (23) and the definition of $\mathbf{C}_P$. The first equality of the second line comes from the bilinearity of the covariance, c.f. Proposition 5.1.2 in (Whittaker, 1990). The next equality follows from (24). Some simple matrix identities and again the definition of the LLSP conclude the proof.

Equivalent formulation of residual using clr transformation

Let us start with the explanatory ratios again i.e. let us show that $\hat{Y}_j(Y)=\hat{Y}_j(Z)$. (For simplicity, we partial on the entire vector Y, but the same argument holds for some $Y_{\mathcal{C}}$.)

$$\hat{Y}_j(Y)=\mathrm{cov}(Y_j,Y)\mathbf{\Sigma}^{-1}Y=\mathrm{cov}(Y_j,\mathbf{F}Z)\mathbf{\Sigma}^{-1}\mathbf{F}Z=\mathrm{cov}(Y_j,Z)\mathbf{F}^T\mathbf{\Sigma}^{-1}\mathbf{F}Z=\mathrm{cov}(Y_j,Z)\mathbf{\Gamma}^{-}Z=\hat{Y}_j(Z). \tag{26}$$

For the first four equalities we were using (6), (3), the bilinearity of covariance (see previous proof) and (5). The last equality we can consider a definition (as in the original definition of the LLSP the covariance matrix needs to be invertible and is here replaced by the pseudoinverse). As this argument was independent of the variable we are predicting, we also conclude that $\hat{Z}_j(Y)=\hat{Z}_j(Z)$.
We now want to show that $Z_j-\hat{Z}_j(Y)=Y_j-\hat{Y}_j(Y)$. Since we have $Y_j=Z_j-Z_D$ and

$$Y_j-\hat{Y}_j=Z_j-Z_D-\widehat{(Z_j-Z_D)}=Z_j-\hat{Z}_j-Z_D+\hat{Z}_D, \tag{27}$$

all we have to show is that $\hat{Z}_D=Z_D$. We have

$$\begin{aligned}\hat{Z}_D(Y) &= \mathrm{cov}\left(\log\frac{X_D}{g(X)},\log\frac{X}{X_D}\right)\mathbf{\Sigma}^{-1}Y=\mathrm{cov}\left(-\log\frac{g(X)}{X_D},\log\frac{X}{X_D}\right)\mathbf{\Sigma}^{-1}Y\\ &= \mathrm{cov}\left(-\frac{1}{D}\sum_{i=1}^{D-1}Y_i,Y\right)\mathbf{\Sigma}^{-1}Y=-\frac{1}{D}\left(\sum_{i=1}^{D-1}\sigma_{il}\right)^T_{l=1,\dots,D-1}\mathbf{\Sigma}^{-1}Y=-\frac{1}{D}\left(\sum_{l=1}^{D-1}\sum_{i=1}^{D-1}\sigma_{il}\sigma_{lk}^{(-1)}\right)^T_{k=1,\dots,D-1}Y\\ &= -\frac{1}{D}\left(\sum_{i=1}^{D-1}\delta_{ik}\right)^T_{k=1,\dots,D-1}Y=-\frac{1}{D}\sum_{k=1}^{D-1}Y_k=-\frac{1}{D}\sum_{k=1}^{D-1}\log\frac{X_k}{X_D}=\log\frac{X_D}{g(X)}=Z_D.\end{aligned} \tag{28}$$

[4] Note that we have $Z=\mathbf{F}^T\mathbf{H}^{-1}Y$. The matrix $\mathbf{F}^T\mathbf{H}^{-1}$ is a right inverse of $\mathbf{F}$.

 ISBN 978-84-947240-2-2

Note that for the first equality of the third line to be true, the geometric mean has to run over the same set as the variables we are partialling on. Here, the fact that we include the reference X_D in the geometric mean only plays a role with respect to the size of the prefactor $1/D$, as $\log(X_D/X_D) = 0$. Together with the independence from the transformation used in the explanatory ratios shown in (26), we conclude that $Y_j - \hat{Y}(Y) = Z_j - \hat{Z}(Z)$.

Pseudoinverse of $\mathbf{\Gamma}$ under permutations of parts in $\mathbf{\Sigma}$

Here we show that a permutation of parts in Y, the variables having covariance $\mathbf{\Sigma}$, corresponds to a simple permutation of the rows in the pseudoinverse $\mathbf{\Gamma}^-$ of the covariance of the corresponding clr-transformed random vector Z. Permuting rows in $\mathbf{\Gamma}^-$ gives

$$\begin{aligned}\mathbf{P\Gamma^-P}^T = \mathbf{PF}^T\mathbf{\Sigma}^{-1}\mathbf{FP}^T = \mathbf{PF}^T\left(\mathbf{Q}_P^T(\mathbf{Q}_P^T)^{-1}\right)\mathbf{\Sigma}^{-1}\left(\mathbf{Q}_P^{-1}\mathbf{Q}_P\right)\mathbf{FP}^T \\ = \mathbf{PF}^T\mathbf{Q}_P^T\left(\mathbf{Q}_P\mathbf{\Sigma}\mathbf{Q}_P^T\right)^{-1}\mathbf{Q}_P\mathbf{FP}^T = \mathbf{PF}^T\mathbf{Q}_P^T\mathbf{\Sigma}_P^{-1}\mathbf{Q}_P\mathbf{FP}^T, \quad (29)\end{aligned}$$

where we applied (5), inserted two identity matrices, and used the identity for permutation of parts in $\mathbf{\Sigma}$ (24). We now observe that

$$\mathbf{F}^T\mathbf{Q}_P^T = \mathbf{F}^T\left(\mathbf{FPF}^T\mathbf{H}^{-1}\right)^T = \mathbf{F}^T(\mathbf{H}^{-1})^T\left(\mathbf{FPF}^T\right)^T = \mathbf{F}^T\mathbf{H}^{-1}\mathbf{FP}^T\mathbf{F}^T, \quad (30)$$

where we used the symmetry of $\mathbf{H}$ for the last equality. This expression contains another auxiliary matrix $\mathbf{G}$ described in (Aitchsion, 1986), which is defined as the $D \times D$ matrix obtained by subtracting from the identity matrix the matrix consisting of elements 1/D. By Properties F1 and G3 in the appendix of Aitchison's book and Properties 5.4 there, we have

$$\begin{aligned}\mathbf{G} &= \mathbf{F}^T\mathbf{H}^{-1}\mathbf{F}, & (31)\\ \mathbf{PGP}^T &= \mathbf{G}, & (32)\\ \mathbf{FG} &= \mathbf{F}. & (33)\end{aligned}$$

With (30) and these properties, (29) simplifies to

$$\mathbf{PF}^T\mathbf{Q}_P^T\mathbf{\Sigma}_P^{-1}\mathbf{Q}_P\mathbf{FP}^T = \mathbf{PGP}^T\mathbf{F}^T\mathbf{\Sigma}_P^{-1}\mathbf{FPGP}^T = \mathbf{GF}^T\mathbf{\Sigma}_P^{-1}\mathbf{FG} = \mathbf{F}^T\mathbf{\Sigma}_P^{-1}\mathbf{F} = \mathbf{\Gamma}_P^-. \quad (34)$$

The last identity follows from (5) again. We have shown that $\mathbf{P\Gamma^-P}^T = \mathbf{\Gamma}_P^-$ through permutation of parts in $\mathbf{\Sigma}$. In Property 5.2 (b) in (Aitchison, 1986) the corresponding identity is stated for $\mathbf{\Gamma}_P$ instead of its pseudoinverse.

Permutation test and local false discovery rates

For the general framework, we follow the procedure outlined in (Hastie et al, 2001). As permuting the residuals seems to lead to a test that is too lenient, we opted for permuting the samples in each column of the original data matrix before evaluating and inverting the covariance matrix. A false-discovery rate (FDR) estimate is obtained for a given cut-off C from dividing the average number of pairs with randomized partial correlations above the cut-off $N_{\mathrm{ra}}(C)$ by the number of pairs with true partial correlations above the cut-off $N_{\mathrm{ob}}(C)$. The number of randomizations was 10,000. All FDRs for cut-offs C in steps of 0.001 were evaluated and a q-value for a given partial correlation value r was determined by taking the minimum over the FDRs corresponding to $C \leq r$. For negative r, the FDRs were determined separately on the other tail of the distribution.

 ISBN 978-84-947240-2-2

WASH your data off: navigating statistical uncertainty in compositional data analysis

F. Ezbakhe[1] and A. Pérez-Foguet[1]
[1]Research Group on Engineering Sciences and Global Development,
Department of Civil and Environmental Engineering, School of Civil Engineering,
Universitat Politècnica de Catalunya BarcelonaTech, Barcelona, Spain
fatine.ezbakhe@upc.edu

Summary

International monitoring of access to drinking water, sanitation and hygiene (WASH) is essential to inform policy planning, implementation and delivery of services. The Joint Monitoring Programme for Water Supply and Sanitation (JMP) is the recognized mechanism for tracking access and progress, and it is based on household surveys and linear regression modelling over time. However, the methods employed have two substantial limitations: they do not address the compositional nature of the data, nor its statistical uncertainty (Ezbakhe & Pérez-Foguet 2018). While the first issue has been tackled previously in the literature (Pérez-Foguet et al. 2017), the effect of non-uniform sampling errors on the regressions remains ignored. This article aims to address these shortcomings in order to produce a more truthful interpretation of JMP data.

The main challenge we try to overcome is how to translate the sampling errors provided in household surveys to the space of compositional data. A Normal distribution is commonly assumed for estimates in household surveys, with a mean and its standard deviation. However, when working with binary data on households – the proportions of households that have access to WASH services – the errors cannot follow normal distributions due to the domain restrictions of proportions, limited to the range 0 to 1. Thus, the Beta distributions seems a better option to characterize the uncertainty around mean access coverage. Yet, as the Beta distribution is defined on the [0,1] interval, the zero values must be dealt with in order to employ the isometric log-ratio (ilr) transformation designed for compositional data. In this article, we investigate the use of two probability distributions (Pearson Type I and Truncated Normal) and Monte Carlo simulations to reinterpret the error in the JMP data so that compositional data analysis is possible.

With a specific focus on the WASH sector, our article shows that the importance of including the survey errors of the data – and its compositional nature – when using this information to support evidence-based policy-making. Indeed, given the current levels of statistical uncertainty in WASH, data may lead to misleading results if errors are not acknowledged (or minimized).

Key words: Demographic Data, Statistical Uncertainty, Compositional Data, Joint Monitoring Programme (JMP), WASH

1 Introduction

In 2015, the global community adopted the 2030 Agenda for Sustainable Development, a universal call to action to end poverty, protect the planet and ensure prosperity to all. The agenda compromises a set of 17 Sustainable Development Goals (SDGs) and 169 targets addressing social, environmental and economic

ISBN 978-84-947240-2-2

aspects of development. To monitor progress towards the SDGs, 232 global indicators are defined and tracked by mandated agencies (UNGA, 2017). The list includes two indicators related to SDG 6.1 and 6.2 targets related to the use of drinking water, sanitation and hygiene (WASH): *(i)* indicator 6.1.1, on the proportion of population using safely managed drinking water services; and *(ii)* indicator 6.1.2, on the proportion of population using safely managed sanitation services, including a hand-washing facility with soap and water.

The task of tracking these two global indicators is undertaken by the WHO/UNICEF's Joint Monitoring Programme for Water Supply and Sanitation (JMP). Since 1990, the JMP has produced national, regional and global estimates of population using improved drinking water sources and sanitation facilities. Specifically, the JMP uses service "ladders" to benchmark and compare across countries (JMP, 2017). For drinking water, the ladder reports on the proportion of the population using: *(i)* drinking water directly from surface water; *(ii)* other unimproved water sources; *(iii)* improved water sources that require more than 30 minutes collection time; *(iv)* improved water sources that require less than 30 minutes collection time; and *(v)* improved water sources that are located on premises, available when needed and free from contamination. Similarly, the ladder for sanitation reports on those with: *(i)* no sanitation at all (open defecation); *(ii)* other unimproved facilities; *(iii)* improved facilities shared between two or more households; *(iv)* improved facilities that are not shared; and *(v)* improved facilities that are not shared with other households and where excreta are safely disposed of in situ or transported and treated off-site.

With this service ladder approach, the JMP generates rural, urban and national estimates for each country, for a total of 26 indicators related to WASH (JMP, 2018). The 8 indicators included in this paper are shown in Table 1. Simple linear regression using ordinary least squares method (OLS) is employed to estimate the proportion of the population using each service level. These estimates are used to monitor progress towards SDG targets, as well as to support informed policy and decision making by national governments, development partners and civil society.

Table 1: 8 primary indicators used by the JMP for monitoring drinking water and sanitation services.

Water	The proportion of the population that uses...
W_1	Piped water drinking water sources
W_2	Other improved drinking water sources
W_3	No drinking water facility (surface water)
W_4	Other unimproved drinking water sources
Sanitation	**The proportion of the population that uses...**
S_1	Improved sanitation facilities connected to sewers
S_2	Other improved sanitation facilities
S_3	No sanitation facilities (open defecation)
S_4	Other unimproved sanitation facilities

However, the "JMP estimation" method has two substantial limitations. First, the compositional nature of the data is not taken into account. The JMP models the service ladder proportions separately, which may derive into untenable results where the sum of the proportions is not equal to 1 (i.e., the whole population). This issue has been addressed previously by Pérez-Foguet et al. (2017), who revealed the importance of considering the compositional nature of WASH coverage estimates for statistical data analysis. Second, the large degree of uncertainty inherent within JMP estimates remains unexplored (Ezbakhe & Pérez-Foguet, Agustí, 2018). This uncertainty stems from sampling errors in the household surveys from which the JMP

draws data and, as such, should be accounted for when estimating WASH coverage.

In this context, and to further support the JMP in the task of improving the modelling of WASH data, this paper investigates how to translate sampling errors provided in household surveys to the space of compositional data.

2 Methodology

In household surveys, a Normal distribution is commonly assumed for estimates, with a mean μ and its standard deviation σ. When working with proportions, however, a Normal distribution is not appropriate, since it may yield values that exceed the 0 and 1 bounds. A Beta distribution is more suitable for the statistical modelling of proportions (Ferrari & Cribari-Neto 2004). Yet, as the Beta distribution is defined on the [0,1] interval, zero values must be dealt with in order to employ log-ratio transformations designed for compositional data.

In this paper, we test the use of two probability distributions to reinterpret JMP data: *(i)* Pearson Type I distribution, a generalization of the Beta distribution bounded to $[\lambda,1-\lambda]$; and *(ii)* Truncated Normal distribution, a generalization of the Normal distribution bounded to $[\lambda_1,\lambda_2]$ (in this case $[\lambda,1-\lambda]$). Their densities are given by Equations 1 and 2, respectively.

$$f(x) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\,(x-\lambda)^{\alpha-1}\,(1-(x-\lambda))^{\beta-1} \tag{1}$$

$$f(x) = \frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})\left(\Phi(\frac{1-\lambda-\mu}{\sigma}) - \Phi(\frac{\lambda-\mu}{\sigma})\right)^{-1} \tag{2}$$

where α and β are the shape parameters of the Pearson Type I distribution, and ϕ and Φ the probability density and cumulative distribution functions of the standard Normal distribution.

The shape parameters are derived from the original data by matching the first and second moments of the "extended Beta" distribution with those of the Normal distribution, as seen in Equations 3 and 4.

$$\mu = \lambda + (1-2\lambda)\frac{\alpha}{\alpha+\beta} \tag{3}$$

$$\sigma^2 = (1-2\lambda)^2\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{4}$$

As suggested by Martín-Fernández et al. (2011), λ can be defined as the "rounding-off error", which relates to the number of significant digits in the database. In this case, we assume $\lambda = 10^{-4}$.

With these two distributions, we use Monte Carlo simulations to generate n sets of JMP data ($n = 1000$). These simulated datasets are used to quantify the uncertainty of JMP data and report the confidence bounds of regressions. For each n simulation, we follow the compositional data (CoDa) methodology: *(i)* we first use a isometric log-ratio (ilr) transformation to bring the compositions to the real space, *(ii)* then apply both ordinary least squares (OLS) linear regression and generalize additive models (GAM) with 4 degrees of freedom to the transformed data, and *(iii)* back-transform the interpolated results to the original scale.

The proposed approach is tested the case of sanitation in rural Madagascar (data in Table 2). The components of the populations are: y_1 sewer, y_2 other improved sanitation facilities, y_3 open defecation, and y_4 other unimproved sanitation facilities. Standard deviations are generated randomly between 0.001 and 0.1, which is the common sampling error in households surveys.

Table 2: JMP data for sanitation in rural Madagascar.

Year	sd	y_1	y_2	y_3	y_4
1992	0.0940	0.0000	0.1300	0.7000	0.1700
1993	0.0720	0.0100	0.2300	0.7300	0.0300
1997	0.0540	0.0000	0.1400	0.7000	0.1600
2000	0.0620	0.0010	0.0865	0.4760	0.4365
2001	0.0600	0.0000	0.0800	0.2600	0.6600
2001	0.0960	0.0000	0.0900	0.2800	0.6300
2002	0.0560	0.0027	0.0887	0.3750	0.5336
2004	0.0730	0.0030	0.0962	0.4620	0.4388
2004	0.0110	0.0012	0.0791	0.5250	0.3947
2005	0.0750	0.0000	0.0909	0.4620	0.4471
2009	0.0040	0.0000	0.0998	0.4910	0.4092
2010	0.0630	0.0360	0.0566	0.5850	0.3224
2011	0.0080	0.0004	0.0843	0.6135	0.3018
2013	0.0540	0.0004	0.1055	0.6052	0.2889
2013	0.0070	0.0010	0.1560	0.5640	0.2790
2016	0.0560	0.0046	0.2317	0.4159	0.3477

3 Results and Discussion

In this section, we compare the coverage estimates obtained by: *(i)* modelling the statistical uncertainty of JMP data with Pearson Type I (aka extended Beta) and Truncated Normal distributions; and *(ii)* applying OLS and GAM regression models.

The importance of translating the sampling errors of JMP data prior to its modelling in the space of compositional data is evidenced in Figure 1. The Normal distribution (Figure 1.a) yields estimates outside the [0,1] interval, specially when proportions of populations are close to the extremes (e.g. in y_1 and y_2). The Pearson Type I distribution (Figure 1.b) may seem suitable to re-interpret the JMP data, as it is delimited at 0 and 1. However, in some cases, it may not be possible to find shape parameters (α and β) that estimate the moments of an Extended Beta distribution. This happens when the mean coverage reported is significantly lower than its standard deviation (e.g. in y_1). Therefore, this approach can only be useful to model uncertainties when standard deviations are smaller than the means. On the contrary, the Truncated Normal distribution (Figure 1.c) is more appropriate to construe the data: it does not only produce estimates between 0 and 1, but also allows for all sets of mean and standard deviation values. Therefore, we choose this latest approach to reproduce the JMP data and model its statistical uncertainty.

On the other hand, when comparing OLS and GAM regressions models (Figure 2), it becomes palpable the need to characterize and represent uncertainty around JMP estimates. In both cases, the 95% confidence interval in the period of JMP data is slightly wide (similar to the errors in the data): (i) with OLS, 0.033, 0.069, 0.097 and 0.105 for y_1,y_2,y_3 and y_4, respectively; and (ii) with GAM, 0.044, 0.089, 0.099 and 0.093. These confidence intervals become much wider in the period beyond the data collected. For instance, in 2020, we can be 95% confident that the expected percentage of the population without access to improved sanitation facilities (aside from open defecation) will be between 40.3% and 61.5% with OLS, or 14.1% and 40.1% with GAM. That it why it is essential to include the survey errors of the data when performing

statistical analysis.

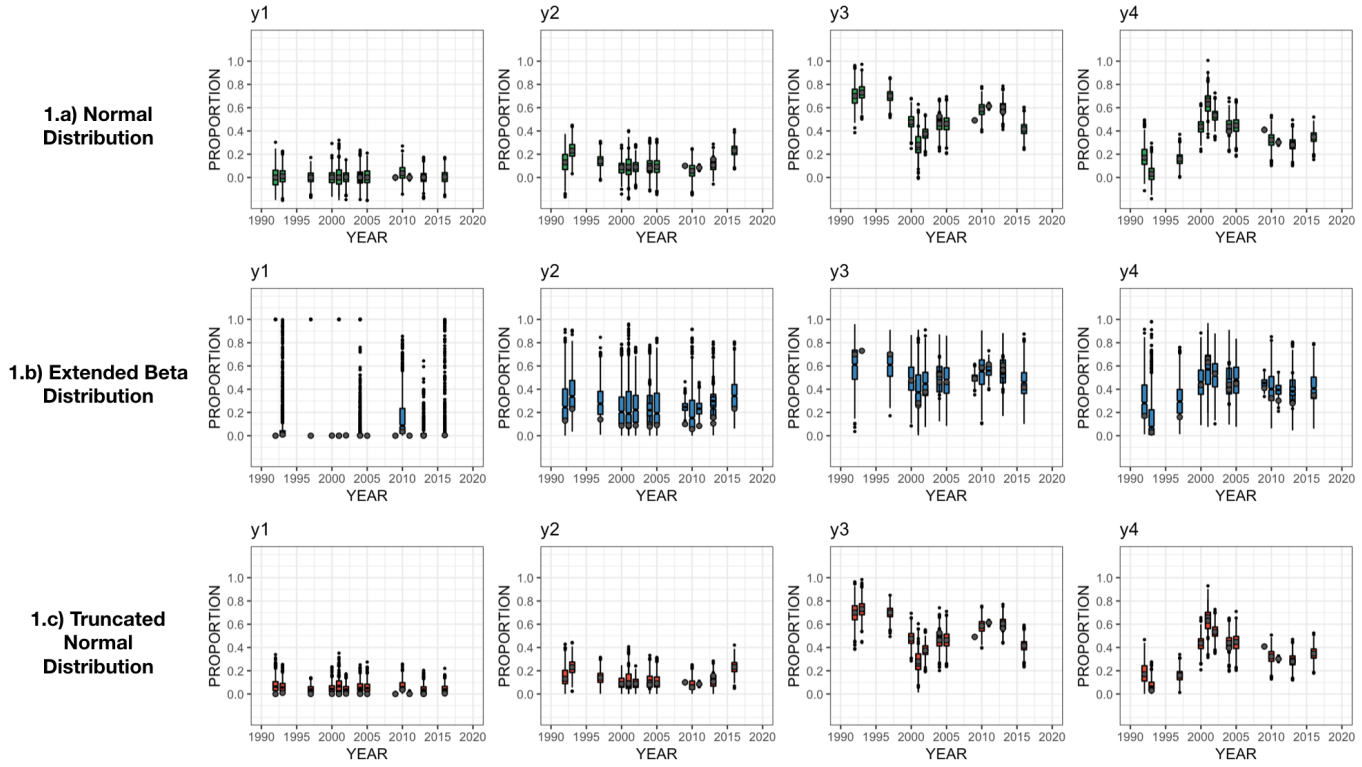


Figure 1: Boxplots of simulated JMP data considering: (1.a) Normal, (1.b) Pearson Type I (aka Extended Beta) and (1.c) Truncated Normal distributions.

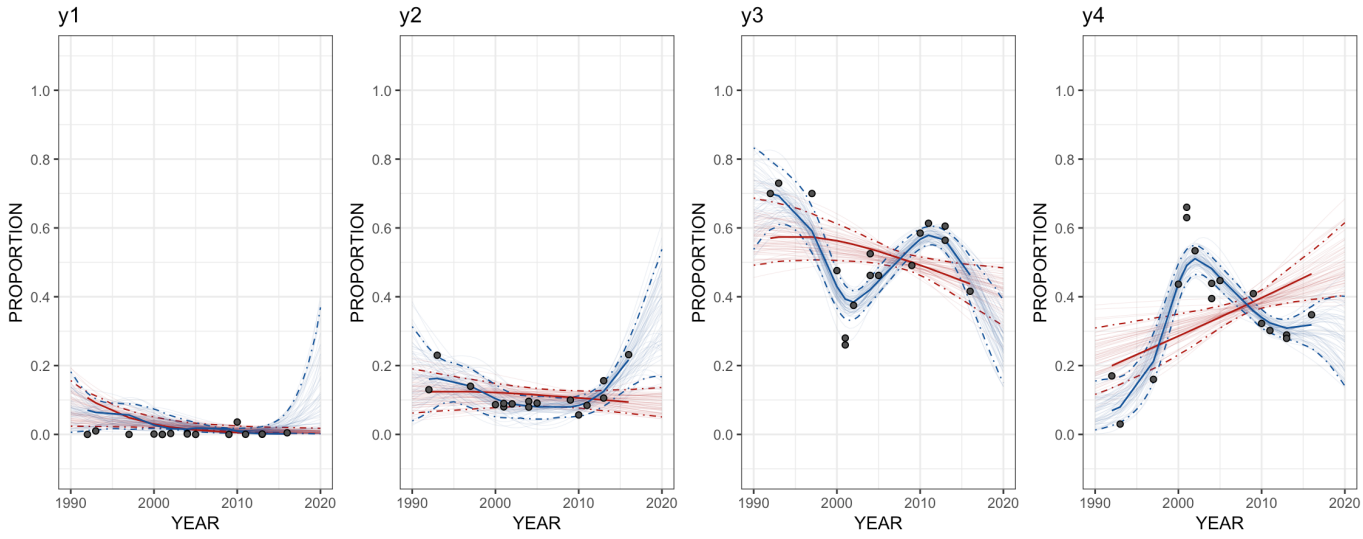


Figure 2: OLS (in red) and GAM (in blue) regressions of JMP data, after ilr-transformation (with 95% confidence intervals using Truncated Normal distributions).

 ISBN 978-84-947240-2-2

Finally, when comparing which regression model is more appropriate, it can be seen that GAM fits better when datasets show nonlinear behaviours. According to the trajectory categorization methodology proposed by Fuller et al. (2016), these components present the following patterns: y_1 "no change" (i.e. the slope for the entire period is close to zero); y_2 and y_3 (i.e. negative but plateauing slope); and y_4 deceleration (positive slope but plateauing below 1). As shown in Table 3, significant improvement is observed when GAM regression is applied to components y_2, y_3 and y_4. Therefore, using GAM results (after ilr transformation) in JMP can lead to more accurate coverage estimates.

Table 3: Values of root-mean-square error (RMSE) for results of models presented in Figure 2.

Model	y_1	y_2	y_3	y_4
OLS	0.0391	0.0522	0.1433	0.1699
GAM	0.0238	0.0238	0.0648	0.0701

References

Ezbakhe, F., Pérez-Foguet, Agustí (2018). Multi-criteria decision analysis under uncertainty: two approaches to incorporating data uncertainty into Water, Sanitation and Hygiene planning. *Water Resources Management 32*(15), pp. 5169–5182.

Ferrari, S., Cribari-Neto, F. (2004). Beta regression for modelling rates and proportions. *Journal of Applied Statistics 31*(7), pp. 799–815.

Fuller, J.A., Goldstick, J., Bartram, J., Eisenberg, J.N.S (2016). Tracking progress towards global drinking water and sanitation targets: A within and among country analysis. *Science of the Total Environment 541*, pp. 857–864.

JMP (2017). Progress on Drinking Water, Sanitation and Hygiene: 2017 Update and SDG Baselines. World Health Organization (WHO) and the United Nations Children's Fund (UNICEF), Geneva.

JMP (2018). JMP Methodology: 2017 Update and SDG baselines. World Health Organization (WHO) and the United Nations Children's Fund (UNICEF), Geneva.

Martín-Fernández, J.A., Palarea-Albaladejo, J., Olea, R.A. Dealing with zeros. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional Data Analysis: Theory and Applications*, pp. 43–58.John Wiley & Sons.

Pérez-Foguet, A., Giné-Garriga, R., Ortego, M.I. (2017). Compositional data for global monitoring: The case of drinking water and sanitation. *Science of the Total Environment 590*,pp. 554–565.

UNGA (2017). Global indicator framework for the Sustainable Development Goals and targets of the 2030 Agenda for Sustainable Development. *Resolution A/RES/71/313*. United Nations General Assembly, New York.

The expression of air quality in urban areas: going further on a Compositional Data Analysis approach

J.Gibergans-Báguena[1], C.Hervada-Sala[1] and E.Jarauta-Bragulat[2]

[1] U. Politècnica de Catalunya - BarcelonaTech (UPC), ESEIAAT, Terrassa, Spain;
jose.gibergans@upc.edu ; carme.hervada@upc.edu

[2] U. Politècnica de Catalunya - BarcelonaTech (UPC), ETSECCPB, Barcelona, Spain;
eusebi.jarauta@upc.edu

Summary

The quality of atmospheric air in large cities is a matter of great importance because of its impact on the environment and on the health of the population. Recently, measures restricting access of private vehicles to the centre of large cities and other measures to prevent atmospheric air pollution are currently topical (Hervada-Sala et al., 2018). The knowledge of air quality acquires special relevance to be able to evaluate the impact of those great social and economic measures.

There are many indices to express air quality. In fact, quite every country has its own, depending on the main pollutants, they have as Plaia and Ruggeri (2011) pointed out. In general, all indices ignore the compositional nature of the concentrations of air pollutants and do not apply methods of Compositional Data Analysis; those indices also have some other weak points such as leak of standardized scale.

A first approach applying Compositional Data Analysis methods has been developed in Jarauta-Bragulat et al., 2016. In the present work, we try to go some step further to improve the understanding and manageability of air quality. The air quality index proposed here takes into account the compositional nature of the data, it has an adequate correlation between input (concentrations) and output (air quality index), it distinguishes between air pollution and air quality and it has a 0−100 reference scale which makes easier interpretation and management of air quality expression. To illustrate the proposed method, an application is made to a series of air pollution data (Barcelona, 2001-2015).

Keywords: Air pollution, Air quality, Air quality index, Health impact, Compositional Data Analysis, Log-ratio.

1. Introduction

Air pollution in cities, mainly in large cities or densely populated areas, is a burning issue that concerns citizens because of its impact on daily life and its consequences on people's health. More than a half of the planet's inhabitants live in urban areas. This is the reason why in large cities the quality of the environment in general and the air in particular is a problem that deserves special attention. Implications of atmospheric air pollution go far

beyond the health of people and affect the environment in general, the economy and the future of life on our planet (Hawking, 2017). Thus, the increasing demand on air quality makes it a point in urban management: cities must take measures to guarantee that air quality is at adequate levels to avoid affections on the health of population. It is very important, therefore, to have an adequate methodology to quantify the expression of atmospheric air quality to help decision makers to control it correctly.

Atmospheric air pollution is usually expressed by a numerical value called "Air Quality Index" (AQI). This index is obtained from concentrations of some air pollutants, usually: O3, CO, NO2, SO2 and suspended particulate matter of certain size or diameter: lower than 2.5 microns (PM2.5) and lower than 10 microns (PM10). Presence of pollutants and particulate matter is expressed by their concentration in units of mass relative to a total volume unit of air usually µg/m3. There are several methodologies to express the quality of atmospheric air from the concentrations of air pollutants. The best known is the proposal by the USA Environmental Protection Agency (EPA), which is based on a piecewise linear function that transforms the concentrations into AQI values in a certain scale. However, all the methodologies applied until now, handle concentrations of air pollutants ignoring their compositional nature and therefore committing some errors, such as calculating arithmetic averages. For more details, see Plaia and Ruggieri 2011 and Jarauta-Bragulat et al. 2016.

The most recent proposal for an Air Quality Index that takes into account the compositional nature of the data is, as far as we know, the one developed in Jarauta-Bragulat et al. 2016. The proposal is based on the concept of logcontrast and an air quality index is defined (AQI*) as a function of the geometric mean of concentrations of six air pollutants (O3, NO2, CO, SO2, PM10, PM2.5). The index is scaled from zero to 100 using a proportionality factor, according to concentration's values.

In the present work, an improvement of that model is proposed, keeping as an essential element taking into account the compositional nature of the air pollutants concentrations, that is, applying Compositional Data Analysis methods. One of the purposes is giving an index that makes it clear that air pollution and air quality mean opposite things, which seems not so clear in most of the existing AQI. At the same time, this index establishes a different slope variation (derivative) in the low pollution zone and the high pollution zone, which allows for a better discrimination in low polluted areas. In addition to a global index of air quality, the proposal is giving an individual index for each of the pollutants, which makes it possible the detection of possible dangerous individual pollutant levels, that otherwise could keep unnoticed in a global index. At last, to help decision makers using it in a reliable, simple and adequate way, the new index has a natural scale to express the values of air quality.

2. Proposal of a new definition of Air Quality Index

Air pollution data are given usually as a real coefficient (N, D+1)-matrix M, where D stands for the number of pollutants and +1 for the period time. Therefore, its k-row can be expressed as:

$$M(k) = \left(t_k \quad c_1(t_k) \quad c_2(t_k) \quad \cdots \quad c_D(t_k)\right),\ k = 1,2,...,N \tag{2.1}$$

being t_k the k-time period (day, week, month, ...), $k = 1,2,...,N$, N the number of time periods and $c_i(t_k)$ the concentration of ith air pollutant at time t_k (units are usually µg/m^3). The most significant air pollutants for their impact on people health in urban surroundings are O3, NO2, PM10 and PM2.5 (Arden-Pope and Dockery 2006; Zhang et al. 2016; Van den Elshout et al. 2014). Therefore, in this work we will consider four air pollutants ($D = 4$) and other components of air grouped in the so-called "residual component". Then, the k-th row we take is

$$M(k) = \left(t_k \quad c_{O3}(t_k) \quad c_{NO2}(t_k) \quad c_{PM10}(t_k) \quad c_{PM2.5}(t_k)\right). \tag{2.2}$$

A logcontrast (LC) is defined as a linear combination of logarithms of parts with coefficients adding up to zero (Aitchison 1986). If a logcontrast is computed considering air pollutants and the residual component, for any time t_k, k=1,2,...,N, put $\alpha_{res} = -1$, then taking into account properties of logarithmic function, the equation we have can be written as:

$$\begin{aligned} &LC = \log\left({c_{O3}}^{\alpha_{O3}} {c_{NO2}}^{\alpha_{NO2}} {c_{PM10}}^{\alpha_{PM10}} {c_{PM2.5}}^{\alpha_{PM2.5}}\right) - \log\left(c_{res}\right) \\ &\alpha_{O3} + \alpha_{NO2} + \alpha_{PM10} + \alpha_{PM2.5} = 1. \end{aligned} \tag{2.3}$$

The filling-up value of air residual component is almost never reported, its computation is very difficult and the second term in the right-hand of Eq(2.3) is almost constant (Jarauta-Bragulat et al., 2016). Therefore, air pollution can be expressed using the considered air pollutants with Eq(2.3) or its approximation

$$\begin{aligned} &f\left(c_{O3}, c_{NO2}, c_{PM10}, c_{PM2.5}\right) = {c_{O3}}^{\alpha_{O3}} {c_{NO2}}^{\alpha_{NO2}} {c_{PM10}}^{\alpha_{PM10}} {c_{PM2.5}}^{\alpha_{PM2.5}} \\ &\alpha_{O3} + \alpha_{NO2} + \alpha_{PM10} + \alpha_{PM2.5} = 1. \end{aligned} \tag{2.4}$$

From Eq(2.4), a function for an Air Pollution Index (API) calculation can be stated as:

$$\begin{aligned} &API = K_{global}\left({c_{O3}}^{\alpha_{O3}} {c_{NO2}}^{\alpha_{NO2}} {c_{PM10}}^{\alpha_{PM10}} {c_{PM2.5}}^{\alpha_{PM2.5}}\right)^{\beta} \\ &K_{global} = \frac{100}{f(O3_{\max}, NO2_{\max}, PM10_{\max}, PM2.5_{\max})} \\ &\alpha_{O3} + \alpha_{NO2} + \alpha_{PM10} + \alpha_{PM2.5} = 1. \end{aligned} \tag{2.5}$$

The exponent β allows applying for a nonlinear model and, thus, having a different shape of the pollution curve in the low pollution zone with respect to the high pollution zone. The value of the exponent used in this work is $\beta = 1.25$, as it has shown to be the best

discriminant. The multiplicative factor K_{global} is introduced for ranging air pollution index values in a scale from zero to 100. The value of the multiplicative factor K_{global} depends on the value of β exponent, as well as on the maximum pollutant' concentrations adopted as maximum admissible pollution (Zhang et al. 2016, "air quality now, 2018"). An example is shown in Table 1 and the corresponding value of K_{global} is 100/223.8.

Once API has been defined, the definition of the Air Quality Index becomes simple:

$$AQI^* = 100 - API \tag{2.6}$$

As a complement to the information provided by the global AQI* values, individual functions for each air pollutant can be calculated as:

$$\begin{aligned}
API_{O3} &= K_{O3}\left(c_{O3}{}^{\alpha_{O3}}\right)^{1.25};\ AQI^*_{O3} = 100 - API_{O3} \\
API_{NO2} &= K_{NO2}\left(c_{NO2}{}^{\alpha_{NO2}}\right)^{1.25};\ AQI^*_{NO2} = 100 - API_{NO2} \\
API_{PM10} &= K_{PM10}\left(c_{PM10}{}^{\alpha_{PM10}}\right)^{1.25};\ AQI^*_{PM10} = 100 - API_{PM10} \\
API_{PM2.5} &= K_{PM2.5}\left(c_{PM2.5}{}^{\alpha_{PM2.5}}\right)^{1.25};\ AQI^*_{PM2.5} = 100 - API_{PM2.5}
\end{aligned} \tag{2.7}$$

	O3	NO2	PM10	PM2.5	*f*	Global AQI*
Scaling factor	0.0	0.0	0.0	0.0	0.0	100
0.10	24.0	40.0	20.0	15.0	22.4	94.4
0.40	96.0	160.0	80.0	60.0	89.5	68.2
0.65	156.0	260.0	130.0	97.5	145.4	41.6
0.85	204.0	340.0	170.0	127.5	190.2	18.4
	240.0	**400.0**	**200.0**	**150.0**	**223.8**	**0.0**
α_i	0.25	0.20	0.30	0.25		

Table 1. Breakpoints for each considered air pollutants; in bold, maximum values assigned. First column indicates the factor to obtain partial breakpoints from corresponding maximum. Column f indicates the result according to Eq(2.4) and exponents are at the last row. In last column, global AQI* values and corresponding colour codes.

3. Case study: Barcelona 2001-2015

A data series 2001-2015 of air pollution in Barcelona is studied. In Figure 1, AQI* functions are plotted for the considered air pollutants. In Figure 2 global AQI* function is plotted. For this period, average of global AQI* values is 90.95, a reasonably good value for air quality. Average individual AQI* values are 91.36 for O3, 93.80 for NO2, 88.40 for PM10

and 86.22 for PM2.5.

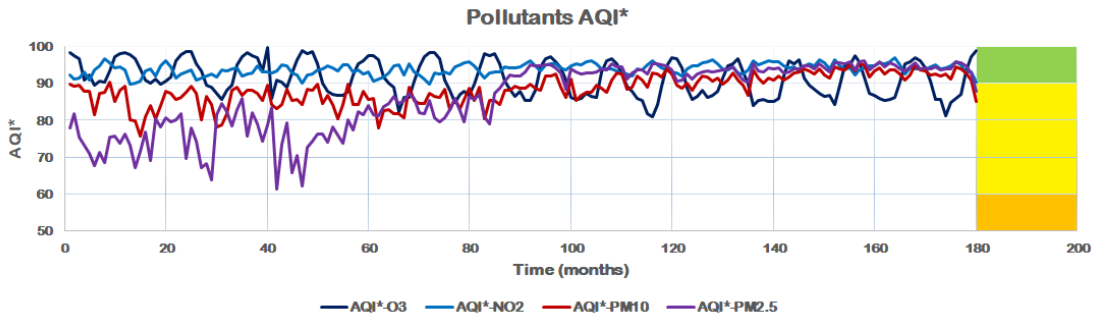


Figure 1. Individual AQI* functions for air pollutants.
Data series: Barcelona 2001-2015.

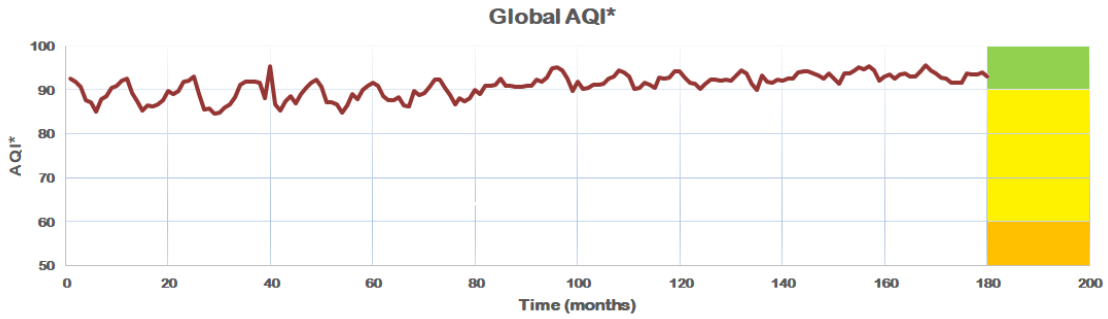


Figure 2. Global AQI* function.
Data series: Barcelona 2001-2015.

4. Conclusions

The main conclusions from this work are:
(1) Definition of an index of atmospheric air quality in cities should be formulated according to the compositional nature of the concentrations and, consequently, applying concepts and methods of Compositional Data Analysis.
(2) The function that should be used in the formulation of an index of air quality is the logcontrast.
(3) The methodology presented in this work allows characterizing air quality individually for each air pollutant, obtaining a global quality index and presenting an AQI-report that illustrates the evolution of global air quality.
(4) Attending to air quality in Barcelona, 2001-2015: (a) Global air quality is satisfactory and has improved in the last ten years. (b) The impact on air quality is mainly due to the presence of particles rather than the presence of gases. (c) In relation to particles, a significant increasing in both corresponding individual quality in last ten years can be

observed. (d) The individual index of ozone has worsened in the last ten years.

Acknowledgements

The authors are grateful to:
(1) The Spanish Ministry of Economy and Competitiveness (MINECO) for the financial support of the research project Fertilecity II "Integrated rooftop greenhouses: energy, waste and CO2 symbiosis with the building. Towards foods security in a circular economy" (CTM2016-75772-C3-1-R; CTM2016-75772-C3-2-R);
(2) The Spanish Ministry of Economy and Competitiveness (MINECO), within the framework of the "CODA-RETOS / TRANSCODA" Project (Ref. MTM2015-65016-C2-1-R and MTM2015-65016-C2-2-R);
(3) The Agency for Management of University Grants and Research (AGAUR) of the Generalitat de Catalunya (GENCAT) within the framework of the project "Analysis of spatial and compositional data" (COSDA, Ref: 2014SGR551, 2014-2016).

References

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

Arden-Pope C, Dockery DC (2006). Health Effects of Fine Particulate Air Pollution: Lines that Connect. Journal of Air & Waste Management Assoc.56:709-742

Hawking S (2017). Stephen Hawking says humans must colonize another planet in 100 years or face extinction. (http://www.cnbc.com/2017/05/05/stephen-hawking-human-extinction-colonize-planet.html)

Hervada-Sala C, Gibergans-Báguena J, Jarauta-Bragulat E (2018). Speed limitation in metropolitan area of Barcelona: impact on air quality. Proceedings of "International Conference on Air Quality", pp 49.

Jarauta-Bragulat E, Hervada-Sala C, Egozcue JJ (2016). Air Quality Index Revisited from a Compositional Point of View. Mathematical Geosciences, 48:581-593.

Plaia A, Ruggieri M (2011). Air quality indices: a review. Rev Environ SciBiotechnol 10:165–179.

Van den Elshout S, Léger K, Heich H (2014). CAQI Common Air Quality Index - Update with PM2.5 and sensitivity analysis. Science of the Total Environment 488–489:461–8. doi.org/10.1016/j.scitotenv.2013.10.060.

Zhang J, Zhang LY, Du M, ZhangW, Huang X, Zhang YQ,Yang YY, Zhang JM, Deng SH, Shen F, Li YW, Xiao H (2016). Identifying the major air pollutants base on factor and cluster analysis, a case study in 74 Chinese cities. Atmospheric Environment 144:37-46.

The Simplicial Generalized Beta distribution - R-package SGB and applications

Monique Graf
Université de Neuchâtel, Switzerland; *monique.p.n.graf@bluewin.ch*
Elpacos Statistics, La Neuveville, Switzerland

Summary

A generalization of the Dirichlet and the scaled Dirichlet distributions is given by the simplicial generalized Beta, SGB (Graf, 2017). In the Dirichlet and the scaled Dirichlet distributions, the shape parameters are modeled with auxiliary variables (Maier, 2015, R-package **DirichletReg**) and Monti et al. (2011), respectively. On the other hand, in the ordinary logistic normal regression, it is the scale composition that is made dependent on auxiliary variables. The modeling of scales seems easier to interpret than the modeling of shapes. Thus in the SGB regression:

- The *scale compositions* are modeled in the same way as for the logistic normal regression, i.e. each auxiliary variable generates $D-1$ parameters, where D is the number of parts.
- The *D Dirichlet shape parameters*, one for each part in the compositions, are estimated as well.
- An additional *overall shape parameter* is introduced in the SGB that proves to have important properties in relation with non essential zeros.
- Use of survey weights is an option.
- Imputation of missing parts is possible.

An application to the United Kingdom Time Use Survey (Gershuny and Sullivan, 2017) shows the power of the method. The R-package **SGB** (Graf, 2019) makes the method accessible to users. See the package vignette for more information and examples.

Key words: Dirichlet distribution, simplicial Generalized Beta, maximum likelihood estimation, imputation, R-package, Time Use survey.

1 SGB distribution

The Dirichlet distribution can be viewed as the distribution of $\mathbf{U} = \mathcal{C}(\mathbf{Y})$, where $\mathbf{Y} = (Y_j)_{j=1,...,D}$ is a vector of independent $Gamma(p_j)$ components and $\mathcal{C}(.)$ is the closure operation (i.e. $U_j = Y_j / \sum_{i=1}^{D} Y_i$). The SGB distribution follows the same construction, with the Gamma distribution replaced by the generalized Gamma, that is the underlying Y_i are independent $GG(a, c\, b_j, p_j)$, c being an arbitrary positive constant. The parameters are all positive and $\mathbf{b} = (b_1, ..., b_D)$ is itself a composition, the *scale composition*. The SGB can also be generated from the Dirichlet:

Definition *Suppose that* $\mathbf{Z} = (Z_1, ..., Z_D)$ *follows a Dirichlet*$(p_1, ..., p_D)$ *distribution. Then the random composition* $\mathbf{U} = (U_1, ..., U_D)$, $(D \geq 2)$, *given by*

$$U_j = \frac{b_j\, Z_j^{1/a}}{\sum_{i=1}^{D} b_i\, Z_i^{1/a}}, \quad j = 1, ..., D \quad or \quad \mathbf{U} = \mathcal{C}[\mathbf{b}\, \mathbf{Z}^{1/a}]$$

ISBN 978-84-947240-2-2

follows a $SGB(a, \{b_j, p_j, j = 1, ..., D\})$ *distribution.*

All parameters are positive; a *is an overall shape parameter,* $\mathbf{b} = (b_1, ..., b_D)$ *a scale composition and* $\mathbf{p} = (p_1, ..., p_D)$ *the vector of Dirichlet shape parameters.*

Conversely, the random composition $\mathbf{Z}$ can be written in function of $\mathbf{U}$,

$$Z_j = \frac{(U_j/b_j)^a}{\sum_{i=1}^{D}(U_i/b_i)^a}, \quad j = 1, ..., D \quad \text{or} \quad \mathbf{Z} = \mathcal{C}[(\mathbf{U}/\mathbf{b})^a]. \tag{1}$$

Because $U_D = 1 - \sum_{j=1}^{D-1} U_j$, there are only $D-1$ variables in the composition $\mathbf{U}$. The L_a-norm of the vector $(\mathbf{u}/\mathbf{b})$ is

$$\|\mathbf{u}/\mathbf{b}\|_a = \left[\sum_{k=1}^{D-1}(u_k/b_k)^a + \left((1 - \sum_{j=1}^{D-1} u_j)/b_D\right)^a\right]^{1/a}.$$

The probability density of the $SGB(a, \{b_j, p_j, j = 1, ..., D\})$ distribution is obtained as

$$f_{\mathbf{U}}(\mathbf{u}_{-D}) = \frac{\Gamma(P)a^{D-1}}{\prod_{j=1}^{D}\Gamma(p_j)} \prod_{k=1}^{D-1}\left\{\frac{u_k/b_k}{\|\mathbf{u}/\mathbf{b}\|_a}\right\}^{ap_k} \left\{\frac{\left(1 - \sum_{j=1}^{D-1} u_j\right)/b_D}{\|\mathbf{u}/\mathbf{b}\|_a}\right\}^{ap_D} \frac{1}{\prod_{k=1}^{D-1} u_k \left(1 - \sum_{j=1}^{D-1} u_j\right)},$$

$$u_k > 0,\ k = 1, ..., D-1, \qquad 1 - \sum_{j=1}^{D-1} u_j > 0.$$

Craiu and Craiu (1969) derived this density. The fitted compositions are defined as the estimated value of the so called Aitchison's expectation $\mathrm{E}_A(\mathbf{U}) = \mathcal{C}[\exp(\mathrm{E}\log(\mathbf{U}))]$, i.e. the image in the simplex of the expectation at log-scale, that is for the SGB, with $\psi(.)$ the digamma function,

$$\mathrm{E}_A(U_k) = \frac{b_k \exp\{\psi(p_k)/a\}}{\sum_{j=1}^{D} b_j \exp\{\psi(p_j)/a\}} \quad k = 1, ..., D.$$

In the R-package **SGB** regression models can be set up for the scale composition $\mathbf{b}$. The shape parameters a and $\mathbf{p}$ are estimated as well, but are supposed constant across compositions.

2 SGB regression model

2.1 Model

The SGB regression models follow the principles of log-ratio analysis advocated by Aitchison (1986). We define a general $D \times (D-1)$ contrast matrix $\mathbf{V}$, such that

$$\mathbf{1}_D^t \mathbf{V} = \mathbf{0}_{D-1}^t,$$

where $\mathbf{1}_D$ is a D-vector of ones and $\mathbf{0}_{D-1}$ is a $(D-1)$-vector of zeros. The model for scales is the general linear model. Let $\mathbf{X}$ be a $n \times p$ matrix of explanatory variables, where n is the sample size. Let $\mathbf{u}_i, i = 1, ..., n$ be the composition associated to $\mathbf{x}_i^t$, the i-th row of $\mathbf{X}$. Then the scales are modeled by

$$\log(\mathbf{b}_i^t)\mathbf{V} = \mathbf{x}_i^t \mathbf{B}, \tag{2}$$

where

$$\mathbf{B} = \left(\boldsymbol{\beta}_1 ... \boldsymbol{\beta}_{D-1}\right)$$

is the $p \times (D-1)$ - matrix of regression parameters for the log-ratio transforms, i.e. the $(D-1)$ columns of $\log(\mathbf{u}_i^t)\,\mathbf{V}$, $i = 1, ..., n$.

2.2 Fitting procedure

There is the possibility to introduce sampling weights into the procedure. These weights $w_i, i = 1, ..., n$ are scaled to sum to n.

The pseudo-log-likelihood is the weighted version of the log-likelihood and is given by

$$\begin{aligned} & \ell\left(a, (b_1, p_1), ..., (b_D, p_D) | \mathbf{u}_{i,-D}, i = 1, ..., n\right) \\ = \; & n\left[(D-1)\log(a) + \log\Gamma(P) - \sum_{k=1}^{D}\log\Gamma(p_k)\right] + \sum_{i=1}^{n} w_i \sum_{k=1}^{D} p_k \log z_k(\mathbf{u}_i) \\ & - \text{ terms not depending on parameters.} \end{aligned}$$

with $z(\mathbf{u}_i) = (z_1(\mathbf{u}_i), ..., z_D(\mathbf{u}_i))$ given at Equation (1) and $P = \sum_{j=1}^{D} p_j$.

The model is estimated by maximizing the pseudo-log-likelihood using a constrained optimization method, the augmented Lagrangian, see e.g. Madsen et al. (2004), and implemented in the R-package **alabama** as function `auglag` (Varadhan, 2015). The gradient is computed analytically and the Hessian numerically. The default constraints are

$$\begin{aligned} a &> 0.1 \quad \text{(to avoid numerical problems)} \\ p_j &> 0, \quad j = 1, ..., D \\ a\,p_j &> \texttt{bound}, \quad \text{by default, } \texttt{bound} = 2.1. \end{aligned}$$

Moments of ratios of parts following the SGB distribution only exist up to $(a\,p_j)$. Thus `bound` $= 2.1$ guarantees the existence of variances of all ratios of parts. Notice that the most important variables, the log-ratios of parts, possess moment of all orders.

A very handy feature of `alabama::auglag` is that the initial values do not need to satisfy the constraints, and that general (twice derivable) constraints on parameters can be introduced. The price to pay is the speed.

3 United Kingdom time use survey 2014-2015

3.1 Data-set

The time diary files of the United Kingdom time use survey 2014-2015 (Gershuny and Sullivan, 2017) provide activities and corresponding level of enjoyment reported over 24 hour period (from 4am to 4am) on business days and week end days. A household file and a individual file contain data collected during the household (individual) interviews. Extrapolation weights at the individual and day levels are given.

In the following application, the total time spent doing an activity during a 24 hour period (`eptime`) was extracted from the diaries. The activities are recorded by 10 min time span. Only the primary activity is considered here. In order to avoid too many zeroes, activities were grouped into 8 categories:

 ISBN 978-84-947240-2-2

`y0` Personal care,
`y12` Employment grouped with study,
`y34` Household and family care grouped with voluntary work,
`y5` Social life and entertainment,
`y6` Sports and outdoor activities,
`y7` Hobbies, games and computing,
`y8` Mass media,
`y9` Travel and unspecified time use.

We consider here the 3,393 (out of 13,603) person-days with zero time spent on `y12`, `y6` and `y7`. For other activities not done that day, the rounded zero technique was used.

3.2 Analysis of one group

The explanatory variables are a weekend indicator; enjoyment data (levels 1 to 7, zero if missing); indicators of missing response on enjoyment; an indicator of "in employment" (`dilodefr=1`); `DVAge` age; `DMSex=2` an indicator of "woman"; indicators of missing `y34`, `y5`, `y8`, `y9`. 16 cases with missing `dilodefr` were deleted. The file with explanatory variables is then `dnot1267b`. The corresponding compositions are in `unot1267b`. The weights are given by `wnot1267b`. The log-ratio transform is alr with reference part `y0` which is never missing. It is specified by the matrix $\mathbf{V}$ = `Vmat2` (see Equation 2). The alr transforms are denoted `a34`, `a5`, `a8` and `a9`. The regression model is specified in the `Formula` `Fnot1267b`, following the syntax of Zeileis and Croissant (2010).

```
Fnot1267b <- Formula(a34 | a5 | a8 | a9 ~
                weekend + enjoy0 +  enjoy34t + enjoy5t +  enjoy8t + enjoy9t +
                I(is.na(enjoy34)) +  I(is.na(enjoy5)) + I(is.na(enjoy8)) +
                I(is.na(enjoy9)) + DVAge + I(DMSex==2) + I(dilodefr==1) +
                ymiss34 + ymiss5  + ymiss8 + ymiss9 )

regnot1267b <- regSGB(Fnot1267b, data = list(dnot1267b, unot1267b, Vmat2),
                weight = wnot1267b, bound = 1.7, shape10 = 0.15,
                control.optim = list(trace = 0, fnscale = -1))

round(table.regSGB (regnot1267b),3)
```

Table 1 shows that the algorithm converged properly to a true maximum of the likelihood: `convergence` equals zero, `kkt1` and `kkt2` (first and second Karush-Kuhn-Tucker conditions) equal one.
`value` is the value of the objective function, minus the pseudo-log-likelihood.
The 78 parameters (`n.par`) are (1+17)*4 regression parameters (intercept and 17 explanatory variables for each alr), one overall shape parameter and 5 Dirichlet shape parameters.
There is the possibility to fix some parameters, but the option was not used (`n.par.fixed` = 0).
`AIC` is Aikaike's criterion.
`Rsquare` was defined by Hijazi and Jernigan (2009) as the ratio of the total variance (Aitchison, 1986) of the fitted compositions to the observed compositions.
`counts.function` and `counts.gradient` give the number of times the objective function and the gradient were evaluated.

More interpretation will be given during the talk.

Table 1: Overall results, output of `table.regSGB`

	statistics
value	-16843.051
n.par	78.000
n.par.fixed	0.000
AIC	33842.101
Rsquare	0.763
convergence	0.000
kkt1	1.000
kkt2	1.000
counts.function	3715.000
counts.gradient	673.000

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman and Hall Ltd (reprinted 2003 with additional material by the Blackburn Press, London (UK).

Craiu, M. and V. Craiu (1969). Repartitia Dirichlet generalizatá. *Analele Universitatii Bucuresti, Mathematicá-Mecanicá 18*, 9–11.

Gershuny, J. and O. Sullivan (2017). *United Kingdom Time Use Survey, 2014-2015. [data collection]*. UK Data Service. http://doi.org/10.5255/UKDA-SN-8128-1.

Graf, M. (2017). A distribution on the simplex of the Generalized Beta type. In J. A. Martín-Fernández (Ed.), *Proceedings CoDaWork 2017*. University of Girona (Spain).

Graf, M. (2019). *SGB: Simplicial Generalized Beta Regression*. R package version 1.0.

Hijazi, R. H. and R. W. Jernigan (2009). Modelling compositional data using Dirichlet regression models. *Journal of Applied Probability & Statistics, 4* (1), 77–91.

Madsen, K., H. Nielsen, and O. Tingleff (2004). Optimization With Constraints. Informatics and Mathematical Modelling, Technical University of Denmark.

Maier, M. J. (2015). *DirichletReg: Dirichlet Regression in R*. R package version 0.6-3.1.

Monti, G. S., G. Mateu-Figueras, and V. Pawlowsky-Glahn (2011). Notes on the scaled Dirichlet distribution. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional data analysis. Theory and applications*. Wiley.

Varadhan, R. (2015). *alabama: Constrained Nonlinear Optimization*. R package version 2015.3-1.

Zeileis, A. and Y. Croissant (2010). Extended model formulas in R: Multiple parts and multiple responses. *Journal of Statistical Software 34* (1), 1–13.

A practical evaluation of the isometric logratio transformation

Michael Greenacre[1], **Eric Grunsky**[2], **John Bacon-Shone**[3]
[1]Universitat Pompeu Fabra, Catalonia; *michael.greenacre@upf.edu*
[2]University of Waterloo, Canada ; *egrunsky@gmail.com*
[3]Hong Kong University, Hong Kong; *johnbs@hku.hk*

Summary

The isometric logratio (ILR) transformation, which is a logratio of geometric means, has been promoted by several authors as the required way, from a theoretical viewpoint, to contrast groups of compositional parts and form a set of new coordinates for analysing a compositional data set. The interpretation of ILRs is made complicated by the fact that each geometric mean depends on the relative values of all the parts included in it. Thus, the geometric mean should never be interpreted as an amalgamation of parts, except in some very special cases that hardly ever occur in practice.

Simple examples can be constructed to show the dangers in using ILRs in statistical modelling. In fact, ILRs should never be used as univariate statistics because of their unclear interpretation. Furthermore, the mathematical properties of the ILR transformation, which justify its existence, are found to be not required for good practice in compositional data analysis.

When groups of parts are required in practical applications, preferably based on substantive knowledge, it is demonstrated that logratios of amalgamations serve as a simpler, more intuitive and more interpretable alternative to ILRs. A reduced set of simple logratios of pairs of parts, possibly involving prescribed amalgamations, is adequate in accounting for the variance in a compositional data set, and highlights which parts are driving the data structure. The necessity to address the research question is also stressed, as opposed to the conversion of the data to ILR coordinates in an automatic way.

Key words: amalgamation, geometric mean, logratio analysis, logratio distance, univariate statistics

1 Introduction

In an empirical study involving compositional data, there is general agreement that ratios of parts constitute the key idea and that logratio transformations conveniently take observed compositions out of the bounded simplex into real space. The simple pairwise logratio is the most easily understood logratio transformation and is the basic building block of John Aitchison's approach to compositional data analysis (CoDa) — Aitchison (1986). For a J-part data set, the set of $J-1$ additive logratios (ALRs) is sufficient to generate every one of the $J(J-1)/2$ pairwise logratios by linear combination. In fact, any set of $J-1$ linearly independent logratios has the same property, and the ratios involved form a network of parts in the form of an acyclic connected graph (Greenacre 2018a, b). The set of J centred logratios (CLRs) is not linearly independent, but forms a very convenient set for computational purposes, since the principal component analysis (PCA) of the CLRs, called logratio analysis (LRA), is equivalent to performing the PCA of the complete set of pairwise logratios — see, for example, Aitchison and Greenacre (2002). Notice that CLRs

ISBN 978-84-947240-2-2

are not intended to be interpreted, nor to be used individually as new variables representing specific parts — they are designed to be used as a complete set to represent the complete set of $J(J-1)/2$ pairwise logratios.

The isometric logratio (ILR) was defined originally by Egozcue, Pawlowsky-Glahn, Mateu-Figueras and Barceló-Vidal (2003) and has seen considerable use since then. An ILR contrasts two subsets of parts, denoted J_1 and J_2, by defining the logratio of their respective geometric means, with a scaling factor:

$$\mathrm{ILR}(J_1, J_2) = \sqrt{\frac{|J_1||J_2|}{|J_1| + |J_2|}} \log \frac{(\prod_{j \in J_1} x_j)^{1/|J_1|}}{(\prod_{j \in J_2} x_j)^{1/|J_2|}} \tag{1}$$

where $|\cdots|$ denotes the number of parts (cardinality), and x_1, x_2, ..., x_J denote the parts. Certain well-chosen sets consisting of $J-1$ ILRs, called "balances", have the property that they form a new orthonormal basis for the compositional data set.

Several authors have adapted the ILR transformation not only as an alternative basis for the logratio space, but also as the definition of new variables that are used in statistical analysis: for example, using them as dependent and independent variables in modelling, making scatterplots of two ILRS and coming to certain conclusions about their relationship, or redefining classical measures involving compositional variables. In this paper we evaluate the validity of using the ILR transformation in practice, focusing especially on its interpretation. In the process we compare the ILR with the much simpler logratio of summed (or amalgamated) parts (SLR):

$$\mathrm{SLR}(J_1, J_2) = \log \frac{\sum_{j \in J_1} x_j}{\sum_{j \in J_2} x_j} \tag{2}$$

Section 2 deals with interpretation issues, showing that ILRs present major problems in practice in understanding what they actually measure. Section 3 lists the claimed benefits of ILRs and gives reasons why these are not necessary prequisites for good practice of compositional data analysis (CoDa). Section 4 concludes with some recommendations.

The overall conclusion is that, while a set of linearly independent ILRs does provide an alternative set of coordinates for a compositional data set, it is a set of uninterpretable coordinates as well as a very complicated way to generate such a set. CLRs, for example, do not pretend to be interpretable variables, they represent the logratio space exactly, are easy to compute and their linear interdependence presents no computational problems. In any case, in almost all studies the research question does not require a new set of variables made up of contrasting groups of parts, so the routine use of ILRs seems artificial and unnecessary. We thus find that ILRs have no significant benefit over simpler transformations, and given the difficulties with their interpretation and the high cost of determining an optimal set of ILRs, their use is preferably avoided. Finally, there is a widespread tendency for them to be casually interpreted as if they are ratios of amalgamated parts, which they are not: hence, we especially warn against the use of single ILRs as summary measures and in modelling.

2 Interpretation of an isometric logratio

The ILR is not a simple transformation and has a complex interpretation. A set of linearly independent ILRs is called a set of "balances", implying a type of contrasting of one "weight" in the numerator with another "weight" in the denominator. The problem is that it is "balancing" geometric means and this is not easy to assimilate.

For example, consider four parts, A, B, C, D, with the following percentages in a larger composition:

A: 10% B: 11% C: 3% D: 4%

Ignoring the scalar normalizing constant in (1), the logratio part of the ILR of $A\&B$ relative to $C\&D$ is $\log(\sqrt{10\times 11}/\sqrt{3\times 4}) = 1.108$, which is positive because the numerator is more than the denominator. The SLR in (2) has no constant, and is simply $\log(21/7) = 1.099$. The two results are similar in this case where the values in the numerator are similar, as well as those in the denominator, so that the geometric means resemble the arithmetic means.

However, suppose now that the values are (with just B changed):

A: 10% B: 1% C: 3% D: 4%

Then $\log(\sqrt{10\times 1}/\sqrt{3\times 4}) = -0.0911$, negative because the numerator is less than the denominator, whereas the SLR is $\log(11/7) = 0.452$, still positive because it has more "weight" in the numerator, which is intuitively correct.

Finally, leaving the same amount of "weight" in the denominator, but redistributing it slightly:

A: 10% B: 1% C: 2% D: 5%

the SLR remains the same but the logratio in the ILR changes to $\log(\sqrt{10\times 1}/\sqrt{2\times 5}) = 0$, which gives the impression that the numerator and the denominator have equal "weight", which they do not.

Clearly, there is a serious intuitive problem if such a statistic is to be called a "balance" and be interpreted as contrasting two groupings of quantities (e.g. percentages of geochemical components, time budgets, money budgets, market shares...). The tendency to give the ILR an interpretation as a ratio of summed values is widespread in the literature, although it is not easy to pinpoint anyone making an unambiguous interpretation of it. The original article on ILRs by Egozcue et al. (2003) makes repeated mention of the ILR transformation *'facilitating the interpretation of the results'* without giving any clear explanation of what the interpretation of an ILR is, apart from it being a *'balancing of groups of parts'*.

Exceptionally, the following explicit interpretation of an ILR was given by Pawlowsky-Glahn, Egozcue and Tolosana-Delgado (2015, page 41):

> *Example 4.7 (Election example).*
> Imagine that in an election, six parties or coalitions have contested. The information collected gives the percentages each party or coalition obtained, that is, a composition in $\mathcal{S}^6$. Divide the parties into two groups, the left and the right wing ones, with four parties in the left wing and two in the right wing. The SBP in Table 4.1 can be used for identifying $\{x_3, x_4\}$ with the right group and $\{x_1, x_2, x_5, x_6\}$ with the left group. If someone is interested in knowing which wing has obtained more votes and in evaluating their relative difference, the balance between the left group versus the right group (first-order partition in Table 4.1) provides this quantitative information:
>
> $$b_1 = \sqrt{\frac{4\cdot 2}{4+2}}\ln\frac{(x_1x_2x_5x_6)^{1/4}}{(x_3x_4)^{1/2}}.$$
>
> The sign of the balance points out which group obtained more votes, and the value gives the size of the difference in log relative scale.

But the above interpretation is incorrect in general. Even the simplest of examples, $x_1 = x_2 = x_5 = x_6 = 15\%$ and $x_3 = x_4 = 20\%$ gives the leftwing a clear majority of 60% to 40% but the ILR, which has the logratio

of log(15/20), is negative. Taking a topical example, at the time of writing the actual election forecasts for the forthcoming Spanish elections give the leftwing coalition's parties percentages of parliamentary seats $x_1 = 36.9\%$, $x_2 = 9.4\%$, $x_5 = 3.4\%$, $x_6 = 2.3\%$, and the rightwing parties $x_3 = 28.0\%$, $x_4 = 20.0\%$ (there are three rightwing parties, so one's forecast has been distributed proportionally across the two others). Using these figures, the leftwing has 52.0% of the seats and the rightwing 48.0%. But the ILR is negative, equal to -1.19 since the geometric mean in the numerator is 7.22 and that of the denominator 23.66. So, if the *'sign of the balance points out which group obtained more votes, and the value gives the size of the difference'*, then the rightwing would win by a huge majority! This demonstrates how small frequencies can radically change the geometric mean.

Buccianti, Nisi and Raco (2015) say that in an ILR the *'ratio measures the relative weight of each group and the logarithm provides the appropriate scale, and a positive balance means that, in (geometric) mean, the group of parts of the numerator has more weight in the composition than the group of the denominator (and conversely for negative balances).'* The qualification *'in (geometric) mean'* raises the problem described above, and the practitioner can easily misinterpret what "weight" in the numerator and denominator actually means. Similarly, Coenders et al. (2015) deal with how tourist expenditures are distributed and compute the ILR of the transportation part relative to the geometric mean of accommodation & food (an amalgamation, no doubt) and activities & shopping (another amalgamation of parts), saying that *'this ratio is used to observe the share of transportation compared to at-destination expenses. Larger values show a higher relative importance of transportation expenses.'* This interpretation is not necessarily valid, based on the examples given previously. Furthermore, they perform statistical tests on this ILR variable, with its unclear interpretation (unless the amounts spent on the two amalgamated categories are similar, in which case the ILR will ressemble the SLR up to a scaling factor).

The book by Pawlowsky-Glahn, Egozcue and Tolosana-Delgado (2015) deals extensively with the theory of ILRs, but includes only simple applications without any interpretation of any specific ILR transformation (except the invalid interpretation pointed out earlier). They use ILRs mainly to reduce the D-part data set to a $D-1$-variable one, and it is the set of ILRs that are used in data exploration and modelling (we return to this aspect in Section 3.1). ILRs are the *'perfect black box'*, as van den Boogaart and Tolosana-Delgado (2013, page 45) say, adding that *'the strongest difficulty with the ilr-transformed values or any orthonormal coordinates [is that] each coordinate might involve many parts (potentially all), which makes it virtually impossible to interpret them in general'*.

There are cases where the ILR transformation is seriously proposed as a univariate statistic, so this is where one can look at how it is interpreted. Buccianti (2015) revises a well-known scatterplot in water chemistry, the Gibbs diagram, which usually has the logarithm of total dissolved solids (an amalgamation) on the vertical axis, substituting this with the logratio that has the geometric mean of the eight dissolved solids as numerator and the amalgamation of all the other components (not their geometric mean) as denominator, which is a type of hybrid of an ILR and a SLR. With this "revision" the approach is stated as now being *'coherent with the nature of compositional data, thus obtaining a simple tool to be used in a statistical sense, going beyond the descriptive approach'* (Buccianti 2015). In fact, the geometric mean of the 8 TDS parts, apart from possible problems of the type described above, is minuscule compared to the other water components, hence the value of this "balance" is, for all practical purposes, almost exactly proportional to log(TDS) used in the classical diagram. The use of the "balance" implies some benefit over the classical Gibbs diagram, but it only adds an unnecessary complication to the plot's interpretation and is not a *'simple tool'* at all.

Morton et al. (2017) also define a single ILR, on a sparse 88×116 data matrix of counts of 116 microbial species in 88 soil samples. They compute an ILR contrasting 86 species with the other 30, i.e. the logratio of two geometric means with 86 and 30 parts respectively. This statistic will be absolutely riddled with the

problem mentioned above. with rare parts radically affecting the values of the geometric means. A much simpler statistic, and more intuitive, would be the logratio of the respective amalgamated parts, where the zeros are no longer problematic. What this ILR of so many parts measures is a mystery, but the authors do suggest that it is contrasting the amount of "stuff" in the numerator with that of the denominator, by showing a diagram of a balance with weights on either side. Pawlowsky-Glahn, Monreal-Pawlowsky and Egozcue (2015, Figure 4) give the same incorrect impression of a balance by drawing it as a physical balance with parts on the left and on the right and saying that *'when the mean balance is placed at the left side, as is the case in Figure 4, it points out that the parts on the left have greater proportions than the parts placed at the right: it works like a lever in equilibrium i.e. a balance in the plain sense.'* Once again, the use of *'greater proportions'* suggests amalgamations of proportions, not geometric means of them. We repeat that the use of an ILR as a univariate statistic is dangerous, because of its unclear substantive interpretation in any application.

To mention one last paradox of using the geometric mean rather than a simple sum in a logratio, suppose that in the time budget of a sample of full-time and part-time workers there are two parts A = percentage of daily time spent at the office working at the desk and B = percentage of daily time at the office drinking coffee, going to the restroom, etc. Suppose that B is consistently a tenth of A, i.e. $B = 0.1A$, a property called *distributional equivalence* in correspondence analysis. Hence, the total time spent at the office is $1.1A$, where A varies across the sample. However, the geometric mean of A and B is $\sqrt{0.1A^2} = 0.316A$ — a researcher studying time budgets of office workers might well be wondering why "time at office" should be measured as $0.316A$ rather than $1.1A$ when used in a logratio contrasting with other daily activities.

3 A set of ILR "balances"

The *raison d'être* of the ILR transformation appears to be in the form of a complete set of ILR "balances", which have the property of being a $(D-1)$-dimensional basis, and thus providing an alternative set of coordinates for the compositional data. The advertised benefits of having this new set of coordinates is that: (i) this ILR basis is orthonormal; (ii) the covariance matrix of this new set of coordinates is nonsingular and thus invertible, when computing Mahalanobis distances or performing multivariate regression (this is often juxtaposed with the centred logratios (CLRs), which are not of full rank); and (iii) they reproduce the logratio geometry (or "Aitchison geometry") exactly. However, none of these three properties are necessary prequisites for good CoDa practice, as we now show.

3.1 Transformation to an orthonormal basis

This supposed benefit is not at all clear. If one wanted an orthonormal basis, the principal components of the CLR-transformed data are the best ones to use and easily computable. Proponents of ILRs say that principal components are not easy to interpret, but then neither are the ILRs: the fact that ILRs have simpler coefficients does not rule out all the paradoxes and complications in their interpretation. In a compositional biplot, the new variables defined by the principal components serve to define the new maximum-variance-explaining axes on which the samples are visualized, supplemented with biplot vectors for the compositional parts (representing the numerators of the CLRs). The coefficients that define the principal components are usually the coordinates of these biplot vectors, although other scalings are possible. If one attempts the same type of analysis with ILRs, the result is unnecessarily complicated by the biplotted vectors now representing the ILRs, with their problematic interpretation.

Trying to emulate a PCA, Martín-Fernández et al. (2018) compute — at considerable expense — a set of "principal balances", and then use the first two as the support axes for a plot of the samples. The problem

here is that they assume the two axes to have some substantive meaning, but their interpretation is unclear and the relationship of the samples back to the original pairwise logratios is lost. Another point to make is that in PCA the component with minimum variance might be of interest since it is an indicator of linear dependence. In the set of "principal balances" the one with minimum variance is inevitably a simple pairwise logratio, which might not even be the logratio with least variance — if such a minimum-variance logratio were of interest it could be found directly by a simple search.

3.2 The nonsingularity of the set of ILR "balances"

Again, this is no benefit all, since it is very simple to cope with a set of singular CLRs representing the data set, inverting it using the generalized inverse. In R, for example, this is achieved by simply using the `ginv` function instead of `solve` in R) if an inverse of the cross-product or covariance matrix is required in a regression analysis or to compute Mahalanobis distances. The multivariate regression and dimension-reduction function `rda` in the **vegan** package accepts sets of singular explanatory variables as a matter of course and with no problem at all. Modern computing makes the advertised nonsingularity of ILR "balances" a redundant property.

3.3 Reproducing the logratio geometry exactly

This is a property that ILR "balances" share with the principal components. However, the insistence on this geometry as some sort of gold standard is perplexing. We know that any data set has a random noise component, so there is no practical reason why one would want to insist on data transformations that represent both the interesting signal and uninteresting noise exactly. This is what makes PCA an interesting method, since it attempts to separate the signal from the noise.

But there are other ways to identify the signal in the multivariate compositions, as described by Greenacre (2018a). One can search for pairwise logratios that maximally explain the variance in the data set, and these logratios can involve amalgamations of parts that are defined by an expert on substantive grounds and for an intentional reason — see Greenacre (2018b) and Greenacre and Grunsky (2019) for further examples, where it is shown in practical applications that a small set of simple pairwise logratios can approximate the logratio geometry very closely, adequate for all practical purposes.

3.4 The substantive value of a set of balances

As a final point, and perhaps the most important of all, if there is any hierarchy or ordering of the compositional parts as a consequence of the type of data or the research question, then logically we should use amalgamations in our logratio modelling, i.e. amalgamation balances (SLRs) of the type (2), not ILRs of the type (1). This means that replacing a compositional data set by a set of ILR "balances" as a general approach to CoDa is clearly inappropriate.

4 Conclusion

ILRs have been promoted as being *'easily interpreted in terms of grouped parts of a composition'* (Pawlowsky-Glahn, Egozcue and Tolosana-Delgado 2015, p. 38). But, as shown here, the geometric mean is a counter-intuitive way of grouping parts and highly sensitive to the presence of rare parts. Since compositional data often have many zeros that have been substituted by small values, these can radically change the geometric means when combined with parts of higher value.

The theoretical benefits of ILRs are not a prequisite for good CoDa practice. If one requires a set of logratios to represent the compositional data set and its logratio geometry exactly, the simply computed CLRs are perfectly adequate for purposes of modelling and dimension-reduction, and they are not intended to have substantive meaning in themselves. Additive logratios can also serve this purpose satisfactorily in many circumstances.

If one requires interpretable logratios that explain a maximum part of variance in the data and closely approximate the logratio geometry, then carefully chosen and substantively meaningful pairwise logratios will be sufficient for all practical purposes (Greenacre 2018a, b). Their choice will depend on the research question: for example, in geochemistry, pairwise logratios that explain differences between sampled regions might well be different from those that correlate with other phenomena of interest, such as environmental variables. In addition, several parts can be amalgamated if this makes sense in the context of the objectives of the research, and these amalgamations can be used in ratios with single parts or other amalgamations (Greenacre and Grunsky 2019). The number of logratios needed to be investigated is relatively few, so this approach is feasible in practice, for as many as 100 compositional parts.

By contrast, it is very difficult to identify the ILR that optimally explains a phenomenon of interest, since there are so many possible groupings of parts, and thus ratios of these groupings. For example, to compute principal "balances" Martín-Fernández et al. (2018) admit that the computations and disk storage required are very high and they only consider data sets up to 15 compositional parts, which are smaller than most data sets in geochemistry and biochemistry. And, even if one identifies the optimal isometric "balance", involving groupings of parts, there still remains the question: what is the "balance" measuring and what does it mean?

Finally, our conclusion here echos what John Aitchison himself said more than 10 years ago: *'When countering this insistence on the use of ILR transformations, I said I would look forward to a convincing practical use of the method. As far as I know there has been no progress in demonstrating its applicability.'* (Aitchison 2008). We have similarly found no convincing demonstration of its applicability, nor any reason why its theoretical properties are a necessary requirement for good CoDa practice. Rather, we find impediments to the use of the ILR transformation and prefer simpler approaches that are directly related to and specifically developed to answer the particular research question.

For a more complete evaluation of the ILR transformation, including empirical examples, see Greenacre and Grunsky (2019). We also highly recommend John Aitchison's writings on this topic — see Aitchison (2008).

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* London: Chapman & Hall. Reprinted by the Blackburn Press.

Aitchison, J. (2008). The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies. *Keynote address, CODAWORK 2008.* URL `https://core.ac.uk/download/pdf/132548276.pdf` (last accessed 16 April 2019)

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

Buccianti, A. (2015). The FOREGS repository: Modelling variability in stream water on a continental scale revising classical diagrams from CoDA (compositional data analysis) perspective. *Journal of Geochemical Exploration* 154, 94–104

Buccianti, A., Nisi, B. and Raco, B. (2015). From univariate background (baseline) values towards

the concept of compositional background (baseline) values. In *Proceedings of the 6th International Workshop on Compositional Data Analysis*, S. Thió-Henestrosa and J.A. Martín Fernández (eds), chapter 5. URL https://upcommons.upc.edu/bitstream/handle/2117/81949/ProceedingsBook.pdf (last accessed 16 April 2019)

Coenders, G., Ferrer-Rosell, B., Mateu-Figueras, G. and Pawlowsky-Glahn, V. (2015) MANOVA of compositional data with a total. In *Proceedings of the 6th International Workshop on Compositional Data Analysis*, S. Thió-Henestrosa and J.A. Martín Fernández (eds), chapter 6. URL https://upcommons.upc.edu/bitstream/handle/2117/81949/ProceedingsBook.pdf (last accessed 16 April 2019)

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

Greenacre, M. (2018a). Variable selection in compositional data analysis, using pairwise logratios. *Mathematical Geosciences*, DOI: 10.1007/s11004-018-9754-x

Greenacre, M. (2018b). *Compositional Data Analysis in Practice.* Boca Raton, Florida: Chapman & Hall / CRC Press.

Greenacre, M. and Grunsky, E. (2019). The isometric logratio transformation in compositional data analysis: a practical evaluation. Submitted to the special issue of *Computers and Geosciences* on "Quantitative understanding of natural phenomena in earth sciences: concepts and tools for data analysis". URL: https://www.researchgate.net/publication/330134599_The_isometric_logratio_transformation_in_compositional_data_analysis_a_practical_evaluation (last accessed 16 April 2019)

Martín-Fernández, J.A., Pawlowsky-Glahn, V., Egozcue, J.J. and Tolosana-Delgado, R. (2018). Advances in principal balances for compositional data. *Mathematical Geosciences* URL https://doi.org/10.1007/s11004-017-9712-z (last accessed 16 April 2019)

Morton, J., Sanders, J.,and Quinn, R.A. et al. (2017). Balance trees reveal microbial niche differentiation. *mSystems* 2 (1) e00162-16. DOI: 10.1128/mSystems.00162-16 (last accessed 16 April 2019)

Pawlowsky-Glahn, V., Egozcue, J.J. and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data.* Chichester: Wiley.

Pawlowsky-Glahn, V., Monreal-Pawlowsky, T., and Egozcue, J.J.(2015). Representation of species composition. In *Proceedings of the 6th International Workshop on Compositional Data Analysis*, S. Thió-Henestrosa and J.A. Martín Fernández (eds), chapter 22. URL https://upcommons.upc.edu/bitstream/handle/2117/81949/ProceedingsBook.pdf (last accessed 16 April 2019)

van den Boogaart, K.G., Tolosana-Delgado, R. (2013). *Analyzing Compositional Data with R.* Berlin: Springer-Verlag.

Multilinear modelling of particle size distributions

K. Khodier[1], M. Lehner[1] and R. Sarc[1]
[1]Montanuniversität, Leoben, Austria; *karim.khodier@unileoben.ac.at*

Summary

Mechanically treating mixed solid wastes, understanding how to influence particle size distributions through shredder parameters can be highly beneficial for concentrating certain materials and improving recovery rates of recyclable materials. Because of the inhomogeneity and variability of these wastes, multilinear empirical modelling is chosen as a practicable approach for describing these influences.

In doing so, the compositional nature of particle size distributions needs to be considered, to obtain valid results when combining the predictions of the models for each dimension. Three potential methods for doing so were identified and analysed a priori for possible restrictions. The application of one of them (modelling the percentage of each particle size, subsequently applying the closure) was further examined with experimental data, using a linear model with two-factor interactions.

It was empirically found, that distortion of the adaption to the calibration points is very high when applying model reduction on each dimension separately. Whereas, when using the same factors and interactions for each dimension, the closure becomes unnecessary, as the summation constraint is fulfilled automatically. The proof of this, as well as the calculation of confidence regions and the comparison with the other presented approaches is subject to further research.

Key words: Municipal solid waste, particle size distribution, multilinear modelling

1 Introduction

Mechanically treating mixed solid wastes, shredders (Fig. 1), followed by screens are often the first machines in the process – generating fractions of liberated particles of defined sizes. Due to differences in the particle size distributions of different materials (Fig. 2), screening also contributes to the concentration of certain materials. The inequality of materials' particle size distributions is present because of differences in original particle sizes as well as comminution behaviour (e.g. brittle fractioning of glass and passing through or tearing of plastic foils). Möllnitz, Khodier et al. (2019) show, that even particle size distributions of different plastic types in the waste are not uniform.

Beyond concentration through screening, particle sizes also influence the process route that the materials undergo – e.g. whether an eddy current separator, which separates non-ferrous metals, is passed. Thus, it desirable to ensure that as much of a certain material as possible passes the machines that are capable of sorting it out.

2 Methods

Considering the described impacts of particle size distributions, understanding how to influence them through parametrisation of the shredder can be highly beneficial for improving recovery rates of recyclable

ISBN 978-84-947240-2-2

Figure 1: Shredder Komptech Terminator 5000 SD.

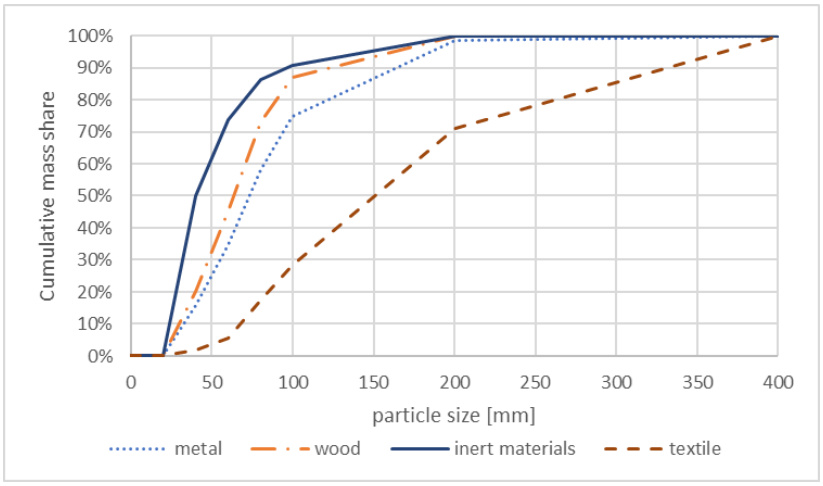


Figure 2: Cumulative particle size distributions of some materials according to Khodier, Viczek et al. (2019)

materials. However, the inhomogeneity and the high variability of waste compositions make it infeasible to do so through detailed physical comminution models. Even if such are found, the necessary information about the material of interest will usually not be available for their practical application. Therefore, it is intended to describe the effects of shredder parameters through multilinear empirical models. The variability of material properties and the resulting comminution behaviour further make it preferable not to stick to a certain kind of model distribution (e.g. Gates-Gaudin-Schuhmann distribution), but using the empirical distribution instead.

Applying multilinear modelling on the empirical distribution of the particle sizes, which is a composition of D particle size classes, the scalar results of the models of each dimension need to be brought together, taking into account their compositional nature. For doing so, the following approaches were identified a priori:

- Modelling the percentage of each particle size, subsequently applying the closure
- Modelling the quantiles of the cumulative distribution at the $D-1$ screen cuts
- Modelling the distribution in isometric log-ratio coordinates, as described by Pawlowsky-Glahn, Egozcue et al. (2015, p. 35)

3 Results

3.1 A priori analysis

Applying these options the following considerations should be kept in mind: Regarding the first approach, applying the closure after the regression distorts adaption to the calibration points. Further a strategy for handling negative model values below 0% or above 100% is needed.

For the second approach it needs to be verified that calibration lines of the quantiles do not cross each other within the model space. Further the model errors for percentages of the finest and coarsest size fractions will differ from the others, as they have fixed contraints at 0% and 100% resprectively, while the both contraints of the other fractions are modelled including uncertainities.

The third approach ensures valid compositions, while needing a handling strategy for zero values – as discussed by Edjabou, Martín-Fernandez et al. (2017) for waste compositions. Furthermore, Khodier and Sarc (2018) point out that the impacts of using the Aitchison-distance in the regression instead of the Euclidian shall be considered. This is especially relevant, as absolute amounts of materials that end in a certain stream are often more relevant for the value added by mechanical waste treatment processes then their proportions.

To find a conclusion about which approach to chose, the results of applying each of them is currently being examined with experimental data. First results are described in the following.

3.2 Empirical findings

Hitherto, the first approach was applied on experimental data with three shredder parameters, using a linear model with two-factor interactions. This model is shown in Equation (1), where f is the resulting mass share of a certain fraction, $\bar{f}$ is the mean value for f over all experimental points, c_i are model constants and A, B and C are the shredder parameters chosen as factors.

$$f = \bar{f} + c_A A + c_B B + c_C C + c_{AB} AB + c_{AC} AC + c_{BC} BC \tag{1}$$

At first, model reduction – eliminating non-significant factors and interactions – was performed for each particle size class separately. As a result, the sum of all fractions was as low as about 70% for for some parameter settings. Therefore, applying the closure would have distorted the calibrated models by almost 50%. Consequently this procedure was discarded.

Thereupon it was examined what happens when using the same factors and interactions for all dimensions. As a result, it turned out, that the closure is unnecessary then, as the fractions automatically sum up to 100%. This is the case, when the sum of the values of $\bar{f}$ for all dimensions is 100% and the sum of the values of each coefficient c_i for all dimensions is 0. The methodology for joint model reduction for all dimensions is still in implementation.

4 Conclusion

Three approaches were identified for handling the compositional nature of particle size distributions when applying multilinear modelling. The approach of modelling the percentage of each particle size was examined using experimental data. Applying model reduction on each dimension separately, it was found that distortion through applying the closure leads to unreliable results, as it gets as high as 50%. Though, when ensuring the use of the same parameters, and therefore of the same model equations, it was found that the closure is not needed anymore, as the constraint that the sum of all fractions must be 100% is fulfilled automatically.

This empirical finding leads to the following hypothesis, that should be proved in the future: Applying multilinear modelling, calibrated through least squares regression, on each dimension of a compositional data set, the resulting model predictions automatically fulfil the summation constraint, as long as it is also fulfilled by all calibration data.

This proof, as well as the examination of confidence regions for the applied approach and the comparison with the other presented approaches is subject to further research.

Acknowledgements

The Competence Center Recycling and Recovery of Waste 4.0 – ReWaste4.0 – (860 884) is funded by BMVIT, BMWFW and the federal province of Styria, whithin COMET – Competence Centers for Excellent Technologies. The COMET programme is administered by FFG.

References

Edjabou, M.E., J.A. Martín-Fernandez, C. Scheutz, and T.F. Astrup (2017). Statistical analysis of solid waste composition data: Arithmetic mean, standard deviation and correlation coefficients. *Waste Management 69*, pp. 13–23.

Khodier, K and R. Sarc (2018). Beschreibung von Abfallzusammensetzungen für Monte-Carlo-Simulationen: Ein Überblick über mathematische Möglichkeiten [Description of waste compositions for Monte Carlo simulations: an overview of mathematical possibilities]. In R. Pomberger (Ed.), *Recy & DepoTech 2018*, pp. 799–804. Leoben: AVAW Eigenverlag.

Khodier, K., S.A. Viczek, A. Curtis, A. Aldrian, P. O'Leary, M. Lehner and R. Sarc (2019). Sampling and analysis of coarsely shredded mixed commercial waste; Part I: procedure, particle size analysis and sorting analysis. *International Journal of Environmental Science and Technology (in submission)*

Möllnitz, S., K. Khodier, R. Pomberger and R. Sarc (2019). Grain Size dependent Distribution of different Plastic Types in coarse-shredded Mixed Commercial and Municipal Waste *Waste Management (in submission)*

Pawlowsky-Glahn, V., J.J. Egozcue and R. Tolosana-Delgado. *Modeling and analysis of compositional data* Statistics in practice. Chichester, West Sussex: John Wiley & Sons Inc.

Statistical evaluation of multi-template PCR biases in microbiome data

Ilia Korvigo[1,2] **, and Evgeny Andronov**[2]
[1]ITMO University, St.Petersburg, Russia *ilia.korvigo@gmail.com*
[2]All-Russia Research Institute for Agricultural Microbiology, St. Petersburg. Russia;

Summary

High-throughput sequencing (HTS) of marker gene libraries has enabled microbiome research on an unprecedented scale and has become the de facto standard in the field. Although the method is renowned for high qualitative phylogenetic sensitivity and accessibility, owing to relative methodological simplicity and low sequencing costs, its quantitative prowess remain an open question in the face statistical properties of the data it generates, namely their compositional nature. Several normalisation and calibration methods have been proposed to both augment marker gene sequencing data with scale and lift compositional constraints, thereby enabling conventional quantitative statistical analysis, though none of them have addressed critical practical issues, such as multi-template PCR bias. Seeing how PCR biases might hamper qualitative microbiome research, we have designed and carried out a study to elucidate temporal dynamics of community compositions undergoing multi-template PCR and, hopefully, provide a prospect of analytical solution in case the problem proves to be too severe to neglect. Our results prove that we cannot carry out reliable quantitative microbiome research without accounting for PCR biases.

Key words: High-throughput sequencing, 16S rRNA, isometric log-ration transform, Bayesian hierarchical modelling

1 Introduction

High-throughput sequencing (HTS) of marker gene libraries has enabled microbiome research on an unprecedented scale and has become the de facto standard in the field. Although the method is renowned for high qualitative phylogenetic sensitivity and accessibility, owing to relative methodological simplicity and low sequencing costs, its quantitative prowess remain an open question in the face of poor reproducibility and, more importantly, statistical properties of the data it generates, namely their compositional nature Gloor et al. (2017). On an abstract level, we can model high-throughout sequencing as a multinomial sampling procedure capturing a relatively small finite number (known as the sequencing depth) of available DNA sequences without altering their distribution in the sample. In microbiome research we map these reads onto a set of distinct community components producing count vectors that are strictly non-negative and are bound by the total sum constraint equal to the sequencing depth. It is easy to see that these counts do not reflect any underlying absolute abundances associated with community components in the sampled environment. Furthermore, any disturbance in individual counts is bound to affect other counts to satisfy the total sum constraint, making independent (orthogonal) reasoning about components nigh-impossible without strong assumptions Gloor et al. (2017). Both properties render conventional statistical tools developed for unconstrained real vector spaces (including all covariance-based methods) inadequate and require a complete overhaul of traditional approaches to quantitative microbiome research. Moreover, there is no way to recover scale (i.e. the information about absolute variations) from the HTS data alone: this problem

ISBN 978-84-947240-2-2

requires additional inputs. Several normalisation and calibration methods based on qPCR and spike-in standards have been proposed to both augment relative abundances (inferred from HTS data) with scale and lift compositional constraints, thereby enabling conventional quantitative statistical analysis Stämmler et al. (2016); Smets et al. (2016). As promising as they sound, none of the papers provide any supporting mathematical evidence to address theoretical contradictions and practical concerns. First of all, spike-in calibration, akin to the additive log-ratio transform (ALR), is not isometric and is sensitive to the choice of component used for calibration. Second of all, HTS itself might not be the only constraining step in an amplicon library preparation workflow. In addition to these theoretical limitations, we must consider one crucial practical issue inherent to the marker gene amplification process. Since distinct DNA templates undergo amplification with a varying degree of efficiency, the process inevitably alters observed component ratios in the amplified mixture Aird et al. (2011); Kalle et al. (2014). Considering that PCR is an exponential growth process, even subtle differences in amplification efficiencies can add up to a significant perturbation of initial unobserved component ratios, undermining not only the spike-in method, but quantitative (and, to a certain extent, qualitative) microbiome research as a whole. Indeed, multi-template PCR is a well recognised source of biases in amplicon libraries, and quite a lot of effort has been put into optimising primers and reaction conditions to minimise distortion Kalle et al. (2014), though we have found no studies on temporal dynamics of community compositions under amplification and not a single publication based on statistical tools appropriate for compositional data, meaning that the actual impact can be both more or less significant than demonstrated in the literature. Seeing how PCR biases might hamper qualitative microbiome research, we have designed and carried out a study to elucidate temporal dynamics of community compositions undergoing multi-template PCR and, hopefully, provide a prospect of analytical solution in case the problem proves to be too severe to neglect.

2 Materials and methods

2.1 Data

Using a qPCR product accumulation assay, we selected 5 PCR cycles (22-26), restricted to the log-linear amplification phase, and sequenced 12 replicates per cycle. All replicates originated from the same faecal matter sample. We also developed a randomisation scheme accounting for possible variance in thermodynamic conditions in our PCR machine. After all data preprocessing steps (including noise elimination) we were left with a sparse count matrix for around 200 unique amplicon sequence variants (ASVs). We removed extremely sparse ASVs and applied a Bayesian-multiplicative zero-replacement strategy with a Dirichlet prior Martín-Fernández et al. (2015).

2.2 Mathematical formulation

We model PCR as a discrete-time process parametrised by initial template counts $\boldsymbol{z} = (z_1, \ldots, z_n)$ and amplification efficiencies. We distinguish amplification efficiencies associated with original DNA sequences extracted from the environment, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, and their amplicons, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)$, due to differences in template lengths, primer-binding site composition and, by extension, primer-template complex stability. This formulation yields the following recurrence for the number of amplicons associated with original template i at cycle t

$$c_i(t) = (\lambda_i + 1) \cdot c_i(t-1) + \theta \cdot z_i$$

 ISBN 978-84-947240-2-2

where $c_i(0) = 0 \quad \forall i$, $\theta_i \in (0, 1]$ and $\lambda_i \in (0, 1]$. This recurrence relation has a corresponding closed-form expression

$$c_i(t) = \frac{\theta_i \cdot z_i}{\lambda_i} \cdot (\lambda_i + 1)^t - \frac{\theta_i \cdot z_i}{\lambda_i}$$

We approximate this equation by discarding the constant term, which is asymptotically irrelevant

$$\hat{c}_i(t) = \frac{\theta_i \cdot z_i}{\lambda_i} \cdot (\lambda_i + 1)^t \sim c_i(t), t \gg 0 \tag{1}$$

Since our model is restricted to the log-linear amplification phase, we can assume unlimited resources and ignore inter-template competition. Consequently, Equation 1 can be trivially extended to a multi-template case

$$\hat{\boldsymbol{c}}(t) = (\hat{c}_1(t), \ldots, \hat{c}_n(t))$$

However, we cannot observe these absolute amplicon counts in HTS data. To get around this issue we use the isometric log-ratio transform (ILR) and model the composition in the space of balances defined by a bipartition strategy and invariant under closure. Although any bipartition strategy will do (because a change of strategy is equivalent to the change of basis in the ILR balance space), the phylogenetic bipartition developed by Silverman et al. Silverman et al. (2017) appears to be a very natural and relatable choice (figure 1). Given a rooted binary phylogenetic tree of n leaves (DNA templates) and $n - 1$ internal nodes,

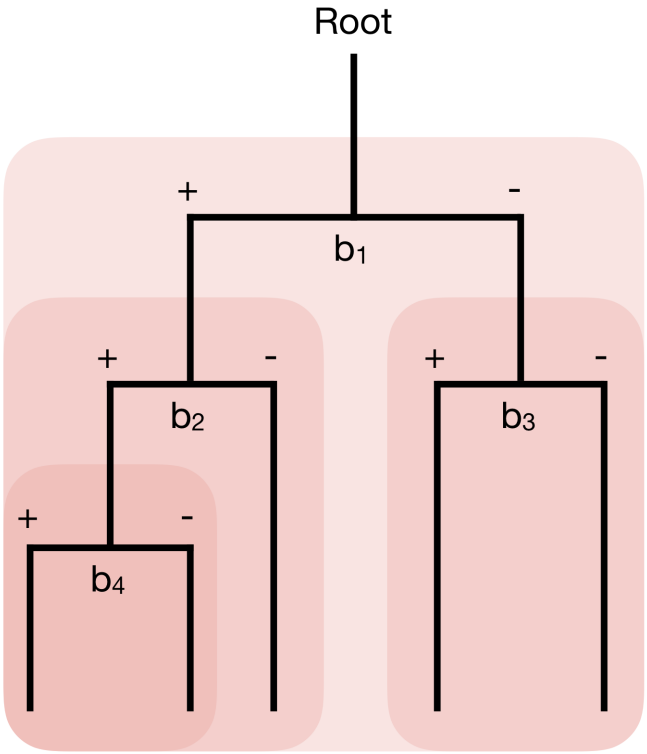


Figure 1: Bipartitions of a rooted binary phylogenetic tree

let's define a sign-matrix Ψ of $n - 1$ rows and n columns such that

$$\psi_{ij} = \begin{cases} -1 & \text{if template } j \text{ belongs to the left subclade of internal node } i \\ +1 & \text{if template } j \text{ belongs to the right subclade of internal node } i \\ 0 & \end{cases}$$

Let's take a closer look at an individual balance

$$b_i(t) = k_i \cdot \log \frac{g(\hat{\boldsymbol{c}}_{\boldsymbol{i}+}(t))}{g(\hat{\boldsymbol{c}}_{\boldsymbol{i}-}(t))}$$

where $g(\hat{\boldsymbol{c}}_{\boldsymbol{i}+}(t))$ and $g(\hat{\boldsymbol{c}}_{\boldsymbol{i}-}(t))$ are geometric means of amplicon counts in the right and left subclades descending from internal node i, $k_i = \sqrt{\frac{n_{i-} \cdot n_{i+}}{n_{i-}+n_{i+}}}$, $n_{i+} = \sum(\psi_{ij} > 0)$, $n_{i-} = \sum(\psi_{ij} < 0)$. We can rearrange the log-ratio of amplicon counts into a sum of three log-ratios

$$b_i(t) = k_i \cdot \left(t \cdot \log \frac{g(\boldsymbol{\lambda}_{i+} + 1)}{g(\boldsymbol{\lambda}_{i-} + 1)} + \log \frac{g(\boldsymbol{\lambda}_{i-})}{g(\boldsymbol{\lambda}_{i+})} + \log \frac{g(\boldsymbol{\theta}_{i+} \cdot \boldsymbol{z}_{i+})}{g(\boldsymbol{\theta}_{i-} \cdot \boldsymbol{z}_{i-})} \right)$$

The equation is a classical linear model for a continuous generalisation over t with a coefficient and intercept defined in terms of the amplification efficiencies and initial template counts. It also allows us to see, that we cannot recover unbiased initial template proportions $\mathcal{C}[\boldsymbol{z}]$, unless $\boldsymbol{\theta} = \boldsymbol{\lambda}$, because there is no free time-dependent parameter associated with $\boldsymbol{z}$. To highlight this this fact, we denote $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$, where $\alpha_i = \log \frac{g(\boldsymbol{\theta}_{i+} \cdot \boldsymbol{z}_{i+})}{g(\boldsymbol{\theta}_{i-} \cdot \boldsymbol{z}_{i-})}$. We used PyMC3 to implement this formulation as a multivariate Gaussian model with a diagonal covariance to infer parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$, assuming a Beta prior for the former, a standard Gaussian prior for the latter and a standard half-Gaussian prior for standard deviations.

3 Results

We used inferred model parameters to predict community composition dynamics from cycle 1 to 30. The community undergoes dramatic changes that can easily rival environmental factors. We visualise the dynamics in terms of phyla proportions (Fig. 3). There are two important conclusions we can draw from our study:

1. any normalisation method designed to reconstruct absolute abundances from marker-gene sequencing data must account for multi-template PCR-biases;
2. while it is technically possible to backtrack distortions associated with amplicon amplification efficiencies, it is impossible to isolate original template amplification biases and their starting populations without additional information and/or strong assumptions, i.e. equality of $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$;

Acknowledgements and appendices

The study was funded by RSF grant 18-16-00073.

References

Aird, D., M. G. Ross, W. S. Chen, M. Danielsson, T. Fennell, C. Russ, D. B. Jaffe, C. Nusbaum, and A. Gnirke (2011). Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology 12*(2), R18.

Gloor, G. B., J. M. Macklaim, V. Pawlowsky-Glahn, and J. J. Egozcue (2017). Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology 8*(NOV), 1–6.

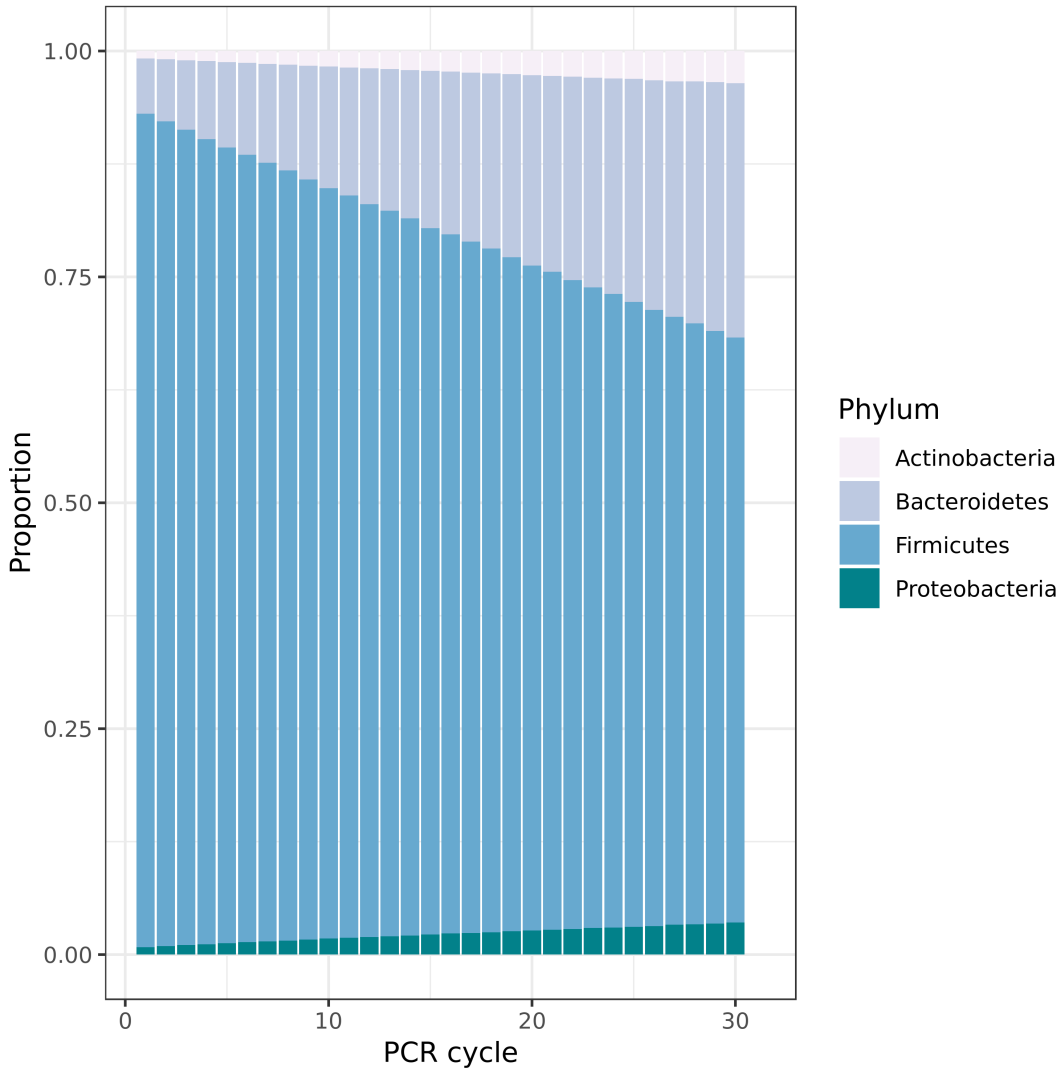


Figure 2: Phyla proportions predicted for cycles 1-30

Kalle, E., M. Kubista, and C. Rensing (2014). Multi-template polymerase chain reaction. *Biomolecular Detection and Quantification 2*(C), 11–29.

Martín-Fernández, J. A., K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo (2015). Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling 15*(2), 134–158.

Silverman, J. D., A. D. Washburne, S. Mukherjee, and L. A. David (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *eLife 6*, 1–20.

Smets, W., J. W. Leff, M. A. Bradford, R. L. McCulley, S. Lebeer, and N. Fierer (2016). A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. *Soil Biology and Biochemistry 96*, 145–151.

Stämmler, F., J. Gläsner, A. Hiergeist, E. Holler, D. Weber, P. J. Oefner, A. Gessner, and R. Spang (2016). Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome 4*(1), 28.

GPAbin for data visualization in the presence of missing observations

N. J. le Roux[1], J. Nienkemper-Swanepoel[1], and S. Lubbe[1]
[1]Stellenbosch University, South Africa; *njlr@sun.ac.za*

Summary

Multiple imputation is a well-established and favored technique for analyzing data containing missing values. It consists of analyzing each imputed data set separately and then combining the estimates for inference. However, the exploratory analysis options of multiple imputed data sets are limited.

Biplots provide a simultaneous configuration of both samples and variables in a two- or three dimensional display. Therefore, a visualization for each of the multiple imputed data sets can be constructed and interpreted individually, but in order to formulate an unbiased conclusion, the multiple visualizations have to be combined for a unified interpretation.

We propose the GPAbin technique to address this challenge for multivariate categorical data sets. The biplots for the multiple imputations are first aligned to a centroid configuration using generalized orthogonal Procrustes analysis (GPA) and then combined by obtaining the mean coordinate matrices from the aligned configurations. The combining step is inspired by Rubin's rules (Rubin, 1987) for combining estimates obtained from analyses applied to multiple imputed data sets. The name GPAbin is derived from the amalgamation of GPA and Rubin's rules.

Simulation studies have confirmed the usefulness of the GPAbin method for categorical data in the context of multiple correspondence analysis (MCA) based biplots. The GPAbin methodology is extended to compositional data containing missing values. This extension replaces MCA based biplots with log-ratio biplots.

The extended GPAbin method is illustrated by creating artificial missingness in a complete compositional example data set. A comparison between the GPAbin- and log-ratio biplots is presented to illustrate the usage of the technique.

Key words: biplots, GPAbin, generalized orthogonal Procrustes analysis, log-ratio biplots, multiple imputation

1 Introduction

Exploratory analysis of categorical data sets is focused on the possible associations between samples regarding responses to variables. Multiple correspondence analysis (MCA) is especially suitable for the simultaneous exploration of multivariate categorical responses in lower dimension. The strength of the association between variables can be investigated and the degree to which the variables are related with respect to the sample responses becomes apparent (Greenacre, 2007). MCA biplots are then constructed to visually explore the results.

Biplots are considered to be a generalization of a scatterplot (Greenacre, 2010) and present each row and column of a matrix as unique vectors, resulting in as many axes as columns. Typically, the rows refer to the samples of a data matrix and the columns to the category levels (CLs) of the variables. The vectors are obtained such that any element of the matrix will be equal to the inner-product of the

ISBN 978-84-947240-2-2

corresponding row and column in the data matrix (Gabriel, 1971). Therefore, biplots can be constructed for multidimensional scaling (MDS) techniques, since observations and responses are displayed according to the interpoint distances between them (Greenacre, 2010). The CLs of categorical variables are illustrated as points referred to as the category level points (CLPs), one point for each CL. The display of the variables, whether it is an axis or CLP, is considered to be the 'framework' or 'scaffolding' of the display (Cox and Cox, 2001). Biplots are low-dimensional representations of high-dimensional data sets in Euclidean space. The low-dimensional display eases visual interpretation and the properties of the Euclidean space enable the use of geometrical properties (Michailidis and De Leeuw, 1998). Since the data points are presented in a reduced dimensional space, the display will always be an approximated representation (Greenacre, 2010). The goal is to utilize a dimension reduction technique that minimizes the amount of information that is lost.

Biplots are not confined to a specific orientation and can therefore be rotated to ease the comparison of multiple displays (Blasius et al., 2009). Orthogonal Procrustes analysis (OPA) allows the comparison of two configurations by using one configuration as the target to which the second (testee) configuration is matched by performing admissible transformations, such as: translation, dilation, reflection and rotation. The optimal configuration is obtained when the sum of squared errors of the distances between the two configurations has been minimized (Ten Berge, 1977; Gower and Dijksterhuis, 2004; Borg and Groenen, 2005). Generalized orthogonal Procrustes analysis (GPA) allows multiple configurations to be matched to a target. The target for GPA is commonly set to the average of the coordinate matrices.

Data containing missing values are a pervasive problem in all data related applications. Multiple imputation (MI) is a preferred unbiased approach to handle missing data (Van Buuren, 2012) where after estimates from each imputed data set can be combined using Rubin's rules (Rubin, 1987) to conduct final inference. There are however limited approaches for exploratory analysis, especially to combine visualizations for multiple imputations. The GPAbin method has been developed to combine visualizations of multiple imputed data sets by first aligning the configurations using GPA and then constructing a final visualization from the mean coordinates of the aligned configurations. GPAbin refers to the use of GPA together with Rubin's rules (Rubin 1987). However, the MI estimates are now regarded as the coordinates of the transformed configurations.

The usefulness of the GPAbin approach has been tested and investigated in an extensive simulation study considering various combinations of percentages of missing values, missing data mechanisms and size of the data matrix. In general, the GPAbin approach performs well under different simulation scenarios, but results in a slightly biased visual representation when the data are simulated from a skewed distribution (e.g. Dirichlet) and the percentage of missing values is high (e.g. 50%). The GPAbin approach achieved the best representation of the complete data simulated from a uniform distribution and also provided unbiased representation for simulations from a normal distribution considering a variety of simulation parameters, as mentioned above.

The GPAbin approach will be extended to constructing log-ratio biplots for compositional data containing missing values.

Section 2 will highlight the steps of the GPAbin approach and Section 3 will show the application of the GPAbin approach to compositional data containing missing values.

2 GPAbin methodology

In general, a MI procedure is applied to the data sets containing missing observations. Configurations of the multiple imputed data sets are constructed and a centroid configuration is obtained from the coordinates. All multiple imputed configurations are aligned to the centroid configuration with GPA. The aligned configurations are then combined into a single configuration using the mean coordinates of the aligned configurations. The methodology is summarized in the flow chart in Figure 1.

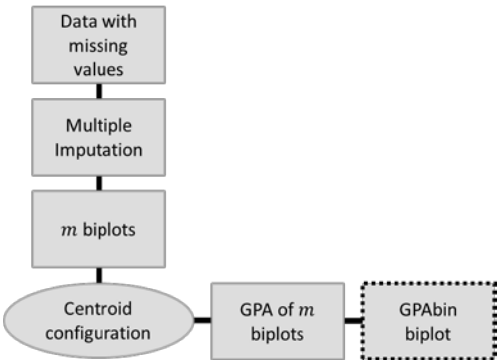


Figure 1: GPAbin methodology.

Considering multivariate categorical data, multiple correspondence analysis (MCA) biplots can be constructed for each multiple imputed data set, where $\hat{\mathbf{X}}_m$ refers to the multiple imputed data sets, with sample coordinates ($\hat{\mathbf{Z}}_m$) and category level points ($\widehat{\mathbf{CLP}}_m$). GPA can be performed on either the samples or CLPs, since the final optimal biplot display can be obtained by transforming the remaining coordinate matrix to align with the optimally rotated coordinate matrix. Here, the target configuration is set as the centroid configuration of the imputed CLPs: $\overline{\widehat{\mathbf{CLP}}} = \sum_{m=1}^{M} \frac{\widehat{\mathbf{CLP}}_m}{M}$.

The GPA procedure minimizes the following: $\min \sum_{m=1}^{M} \left\| s_m(\widehat{\mathbf{CLP}}_m)\mathbf{Q}_m - \overline{\widehat{\mathbf{CLP}}} \right\|^2$, where s_m is the optimal dilation factor for the m^{th} imputation, $\widehat{\mathbf{CLP}}_m$ is the coordinate matrix for the CLPs of the m^{th} imputation, $\mathbf{Q}_m$ is the optimal orthogonal rotation matrix for the m^{th} imputation and $\overline{\widehat{\mathbf{CLP}}}$ is the coordinate matrix for the centroid configuration of the imputed CLPs. The translation of the GPA is incorporated by centering the configurations at the origin before comparing the configurations (Gower and Dijksterhuis, 2004). The transformed coordinate matrices for each imputation are obtained as follows: $\widehat{\mathbf{CLP}}_m^* = s_m(\widehat{\mathbf{CLP}}_m)\mathbf{Q}_m$. The transformed biplot of each imputed data configuration can now be constructed by applying the same dilation and rotation factors on the samples as is used for the CLPs: $\hat{\mathbf{Z}}^* = s_m\hat{\mathbf{Z}}_m\mathbf{Q}_m$.

The final combined configuration is obtained by applying a variation of Rubin's rules by constructing a biplot from the mean coordinates of the samples and CLPs, resulting in a GPAbin biplot with the following coordinate matrices: $\overline{\widehat{\mathbf{CLP}}}^* = \sum_{m=1}^{M} \frac{\widehat{\mathbf{CLP}}_m^*}{M}$ and $\hat{\bar{\mathbf{Z}}}^* = \sum_{m=1}^{M} \frac{\hat{\mathbf{Z}}_m^*}{M}$.

As previously stated, this method is referred to as GPAbin, indicating the use of GPA to optimally rotate MI configurations toward a centroid (target) configuration and subsequently applying Rubin's rules to obtain the mean coordinates of the transformed CLPs ($\widehat{\mathbf{CLP}}_m^*$)and samples ($\hat{\mathbf{Z}}_m^*$) for a final display. Alternatively, the GPAbin procedure can be simplified by using the coordinate matrices of the configurations without distinguishing between sample coordinates and CLPs. The steps of the GPAbin procedure will be unchanged.

3 Application

The application of the GPAbin method on a compositional data set will be illustrated using the *cups* data

ISBN 978-84-947240-2-2

set available in the *easyCODA* R package (Greenacre, 2018). The data set consists of 47 samples, Roman cups, and the compositions of eleven oxides observed in each cup. All biplots will consist of green triangular plotting characters which represent the oxides, and grey circular plotting characters which represent the samples (cups). The log-ratio biplot of the complete data set is presented in Figure 2.

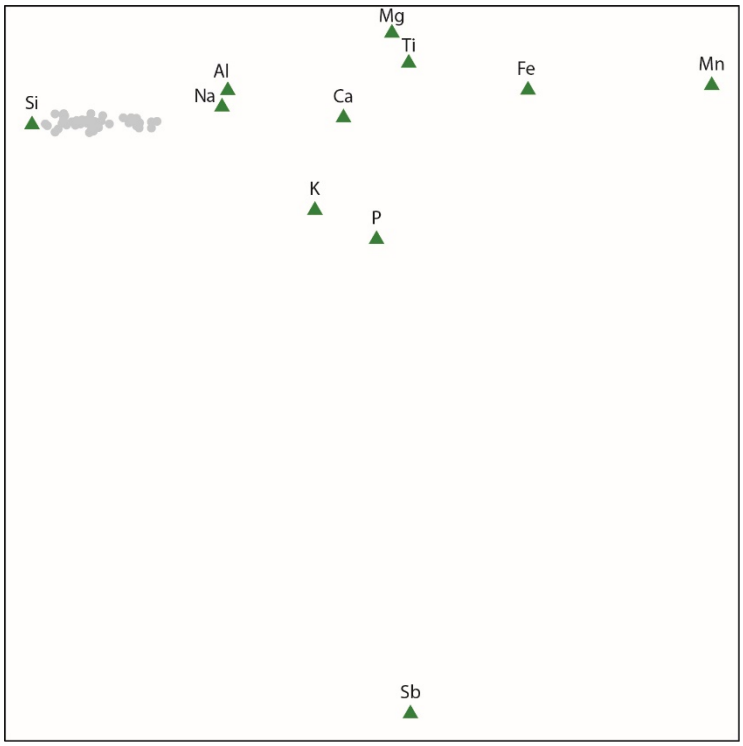


Figure 2: Log-ratio biplot of complete *cups* data set.

The interpretation of a log-ratio biplot relies on the link vectors between variable coordinates that represent pairwise log-ratios and does not focus on the positions of the coordinates in the biplot display (Greenacre, 2010). This differs from MCA biplots where the positions of the sample coordinates and CLPs express the strength of the associations between samples and certain responses.

Missing values are inserted with a missing completely at random (MCAR) missing data mechanism (MDM), which means that all missing values are independent of the observed and other unobserved observations. Different percentages of missing values were explored, but only the 30% missing value figures will be presented here. The missing data are imputed five times using the *mice* R package (Van Buuren and Groothuis-oudshoorn, 2011). The imputed data sets are closed before constructing the log-ratio biplots for each completed data set presented in Figure 3.

The MI log-ratio biplots are optimally aligned to the GPA centroid configuration, which is the mean coordinates of the five log-ratio biplots presented in Figure 3. After the imputed log-ratio biplots are aligned, the mean coordinates are calculated and used to construct the GPAbin biplot presented in Figure 4.

ISBN 978-84-947240-2-2

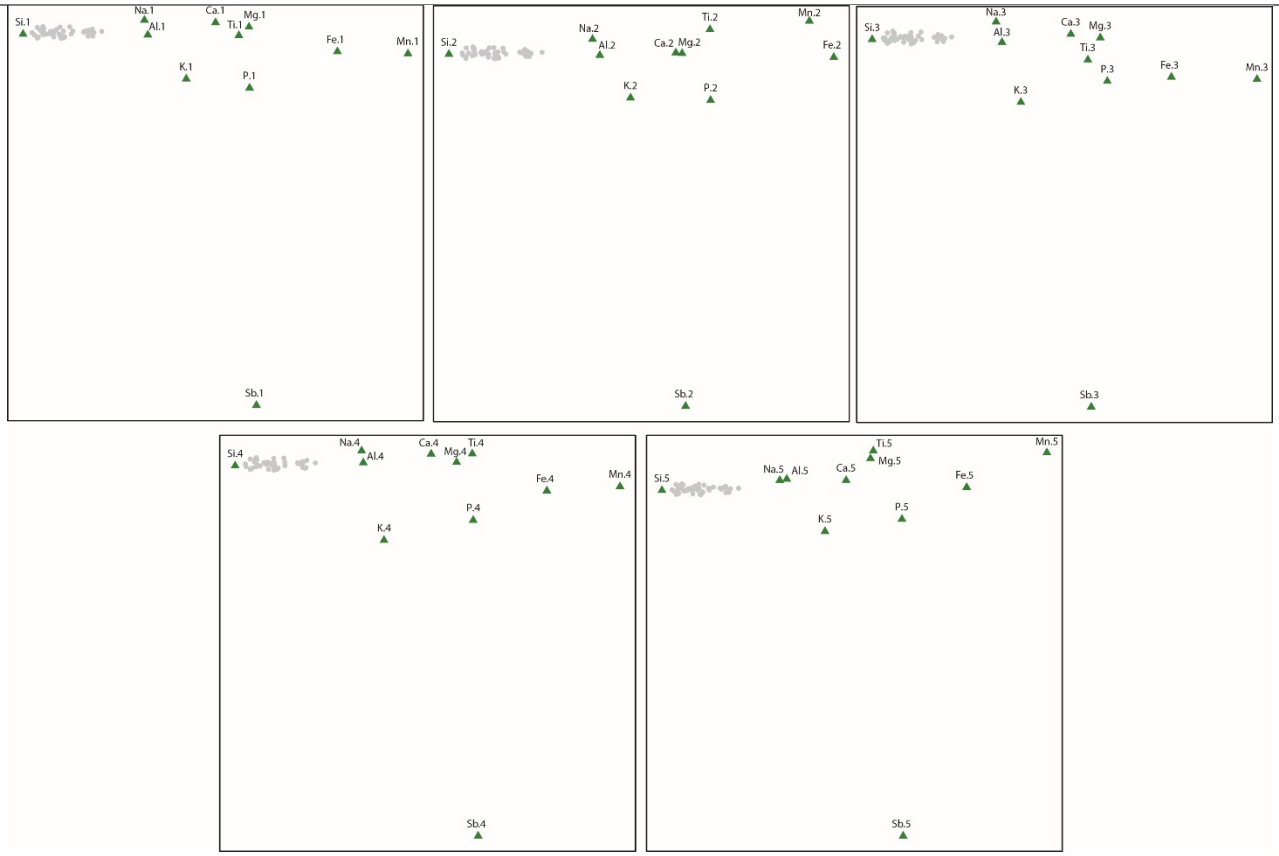


Figure 3: Log-ratio biplots of five multiple imputations of *cups* data set using *mice*.

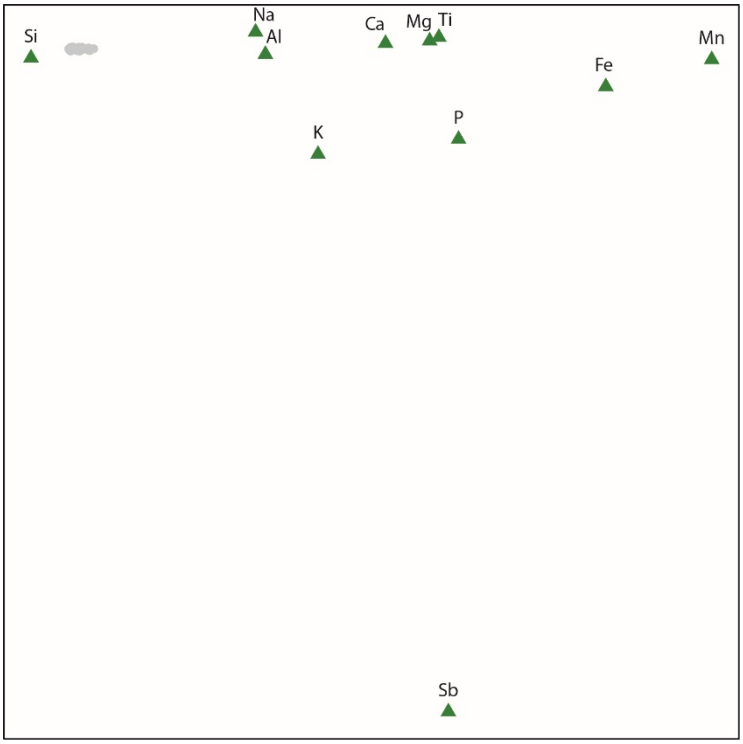


Figure 4: GPAbin biplot of five imputations.

ISBN 978-84-947240-2-2

The GPAbin configuration (Figure 4) is compared to the log-ratio biplot (Figure 2) using OPA which enables the calculation of two measures of fit: Procrustes statistic (0.09) and congruence coefficient (0.97). Both measures of fit considers the configurations before the application of OPA. The Procrustes statistic expresses the magnitude of the admissible transformations (translation, dilation, rotation and reflection) to be performed on the testee in order to match the target configuration. A value between zero and one is calculated, with a smaller value indicating good fit. The congruence coefficient, also a value between zero and one, compares the distances between the coordinates in the target and testee configurations before the application of OPA and is interpreted as a measure of determination with a value close to one indicating good fit.

The bias is determined using three measures: absolute mean bias (0.42), root mean squared bias (0.94) and mean bias (0.12). All resulting in bias measures close to zero.

The measures of comparison confirm that the GPAbin biplot (Figure 4) preserves the configurations observed in the log-ratio biplot (Figure 2) between the oxides and the cups as were observed from the visualizations.

4 Concluding remarks

This paper showed the extension of the GPAbin method, developed for multivariate categorical data, to compositional data. The GPAbin method provides a solution for visualization of data containing missing values when MI techniques are used to complete the data. The GPAbin method has shown to successfully preserve the configurations in the visualizations of multiple imputed data sets when compared to the original complete visualization. The GPAbin method also results in unbiased representation of the complete visualization approximation obtained in two dimensions.

Future considerations would be to apply other MI procedures to determine which are suitable for compositional data analysis for GPAbin visualization. In this application the *mice* algorithm performed well, as illustrated in the MI biplots (Figure 3) which resulted in similar representation to the complete log-ratio biplot in Figure 2. The GPAbin method could also be further extended by considering coordinate transformation methods by Pawlowsky-Glahn and Buccianti (2011).

The GPAbin biplot enables a global visual representation of multiple imputed data sets which unifies visual interpretation, which can be further scrutinized for a variety of biplot applications.

References

Blasius, J., Eilers, P. H. C. and Gower, J. (2009). Better biplots. *Computational Statistics and Data Analysis*. Elsevier B.V., 53(8), pp. 3145–3158.

Borg, I. and Groenen, P. (2005). *Modern Multidimensional Scaling*. 2nd edn. United States of America: Springer.

Cox, T. and Cox, M. (2001). *Multidimensional Scaling*. 2nd edn. Boca Raton: Chapman & Hall/CRC.

Gabriel, R. K. (1971). The biplot graphic display of matrices with application to principal component analysis, *Biometrika*, 58(3), pp. 453–467.

Gower, J. C. and Dijksterhuis, G. B. (2004). *Procrustes Problems*. Oxford: Oxford University Press.

Greenacre (2018). *Compositional Data Analysis in Practice*. Chapman & Hall/CRC Press.

Greenacre, M. (2007). *Correspondence Analysis in Practice*. 2nd edn. Boca Raton: Chapman & Hall/CRC.

Greenacre, M. (2010). *Biplots in Practice*. Fundación BBVA.

Michailidis, G. and De Leeuw, J. (1998). The Gifi System of Descriptive Multivariate Analysis, *Statistical Science*, 13(4), pp. 307–336.

Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional Data Analysis Theory and Applciations*. United Kingdom: John Wiley & Sons Ltd.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Ten Berge, J. M. F. (1977). Orthogonal procrustes rotation for two or more matrices, *Psychometrika*. Springer-Verlag, 42(2), pp. 267–276.

Van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Boca Raton: Chapman & Hall/CRC Interdisciplinary Statistics Series.

Van Buuren, S. and Groothuis-oudshoorn, K. (2011). mice : Multivariate Imputation by Chained, *J Stat Softw*, 45(3).

ISBN 978-84-947240-2-2

Compositional Data Analysis of Finnish Birch Sap

T. Lillhonga[1] and H. Kallio[2]

[1]Novia University of Applied Sciences, Vasa, Finland;
tom.lillhonga@novia.fi

[2]Food Chemistry and Food Development, University of Turku, Finland

Abstract

Sap is a nutrient solution produced during the spring season on the northern latitudes by some of the deciduous trees, like birch and maple species. Birch sap is a clear and dilute water solution of fructose, glucose, organic and phosphoric acids, proteins, and metal cations. The slightly sweet sap can be consumed as it is or used as raw material in manufacturing of other health-beneficial products, in cosmetics, or to replace water in beer production etc. There is a growing interest on the world market for products based on birch sap (Seal 2016), motivating in depth studies of the properties and chemistry of birch sap.

Sap samples from Finnish birch trees (Beteula pubescens and Betula pendula) were collected during the spring seasons of 2016-18. Samples were collected daily from each tree during the three seasons. The same, selected trees were sampled each year, to minimize the between-tree variation. A total of 240 samples were collected during the three years and each sample was analyzed for monosaccharides, metal cations, pH and conductivity. Birch sap has previously been studied in a number of publications (Kallio et al. 1987), but more in a univariate context. Our acquired data matrix is compositional (CoDa) and multivariate to its nature and therefore analyzed with appropriate multivariate CoDa-methods. Compositional statistics and CoDa-PCA models are calculated to study the evolvement of the time series for the different tree species. The acquired results shows interesting variations and correlations over time and sheds new light on important questions concerning sap quality and sap composition.

Key words: birch sap, seasonal variation, multivariate, compositional data

1 Introduction

Birch trees (Beteula pubescens and Betula pendula) are deciduous and widely spread trees in the Nordic countries. They show the property of "bleeding" during the spring season, producing a clear, water like, liquid called birch sap. The sap is produced by the roots and transported through the tree trunk up to the branches supporting the incipient blooming. Sap consists mainly of water, but contains also sugars, metal ions, organic acids and amino acids, all vital nutrients for the tree. People in the Nordic countries, especially in Finland and Sweden, have for centuries been collecting birch sap by drilling small holes in the tree trunk letting the sap pour into vessels or bottles. Sap is considered to have health beneficial properties. The global health industry has shown increasing interest in sap and other similar organic

drinks, like i.e. coconut water. Sap is also used as raw material in the cosmetics industry and as raw material for food and beverages, i.e. production of birch sap beer. Birch sap is an interesting raw material for microbreweries in the Nordic countries enhancing the use of local raw materials and for branding purposes. Today birch sap is collected in industrial scale and it is important to better understand the chemical content of the sap and its variations during the sap period. Many studies have studied and described the concentrations of sugars (Kallio and Ahtonen (1); 1987), metal cations (Huldén and Harju, 1986), organic acids (Kallio and Ahtonen (2), 1987) and amino acids (Ahtonen and Kallio, 1989) in sap. All previous studies have used traditional univariate statistics comparing concentrations of different species and calculation correlations. Our focus is to study the main contents of birch sap and their seasonal variations using compositional methods. The differences or similarities between the samples can be found in studying concentration ratios rather than comparing absolute values.

2 Materials and Methods

Sap samples were collected during three consecutive sap seasons 2016-2018 from different trees in the Ostrobothnia region, in the western costal part of Finland. During spring 2016 we collected sap from six different trees, 2017 from ten different trees and during 2018 from eight different trees. Four of the trees were sampled all three years, four trees were sampled during 2017-18 and three trees were sampled only one year. Every collected sample was analyzed for glucose, fructose (ppm) and the elements K, Ca, Mg, Zn, P and Mn (mg L^{-1}). In addition, pH and electrical conductivity (μS cm^{-1}) were measure for all samples. The sugars were measured with an IC(PAD) instrument, the elements Ca and K with an AAS instrument and the rest of the elements with an ICP/MS instrument. Several other trace metal elements were also measured with the ICP/MS instrument but their concentrations were in general below the detection limit and hence omitted from further analysis. The concentration ranges and median values are presented in Table 1.

Table 1: Ranges and median values of chemical and physical parameters measured in birch sap.

	Glucose (ppm)	Fructose (ppm)	K (mg/L)	Ca (mg/L)	Mg (mg/L)	P (mg/L)	Mn (mg/L)	Zn (mg/L)	pH	Conduct. (μS cm^{-1})
Range	46-12175	910-11480	20.7-207.8	14.6-137.2	3.9-44.3	0.5-28.7	0.6-22.4	0.3-5.0	5.2-7.3	203-1998
Median	2989	3846	79.3	50.3	14.7	8.7	2.7	1.4	6.3	521

A total of 240 samples were analyzed, 43 during year 2016, 129 during 2017 and 68 samples during 2018. Samples are identified by the initials of the forest owner, followed by a day-number representing the date the sample was collected. All three seasons lasted between 8th of April to the 6th of May. The 8th of April, being the first day, labeled day 1 and the last day, 6th of May labeled day 29. The sample periods were of different lengths. The first season during 2016 lasted from 13th of April (day 6) to 21th of April (day 14), the second season 2017 started 8th of April and lasted until 6th of May (day 29) and the last season 2018

started 19th of April (day 20) and lasted until 28th of April (day 21). Samples were not taken every day from all trees and thus the day-numbers helps following the time series between different years. All calculations were executed in the software R (R Core Team, 2018) and package robCompositions (Templ, Hron et al., 2011).

3 Results and Discussion

Compositional Principal Components Analysis (PCA) models were calculated separately for each sap season and a combined PCA-model for all samples. The data is first expressed in isometric logratio coordinates, followed by a robust PCA calculation. Finally, the scores and loadings are back-transformed to the clr-space in order to visualize the biplot (Filzmoser). The use of robust estimators makes the model less sensitive to potential outliers. The first model is calculated on the 43 samples from sap season 2016. The result is presented as a biplot in Figure 1.

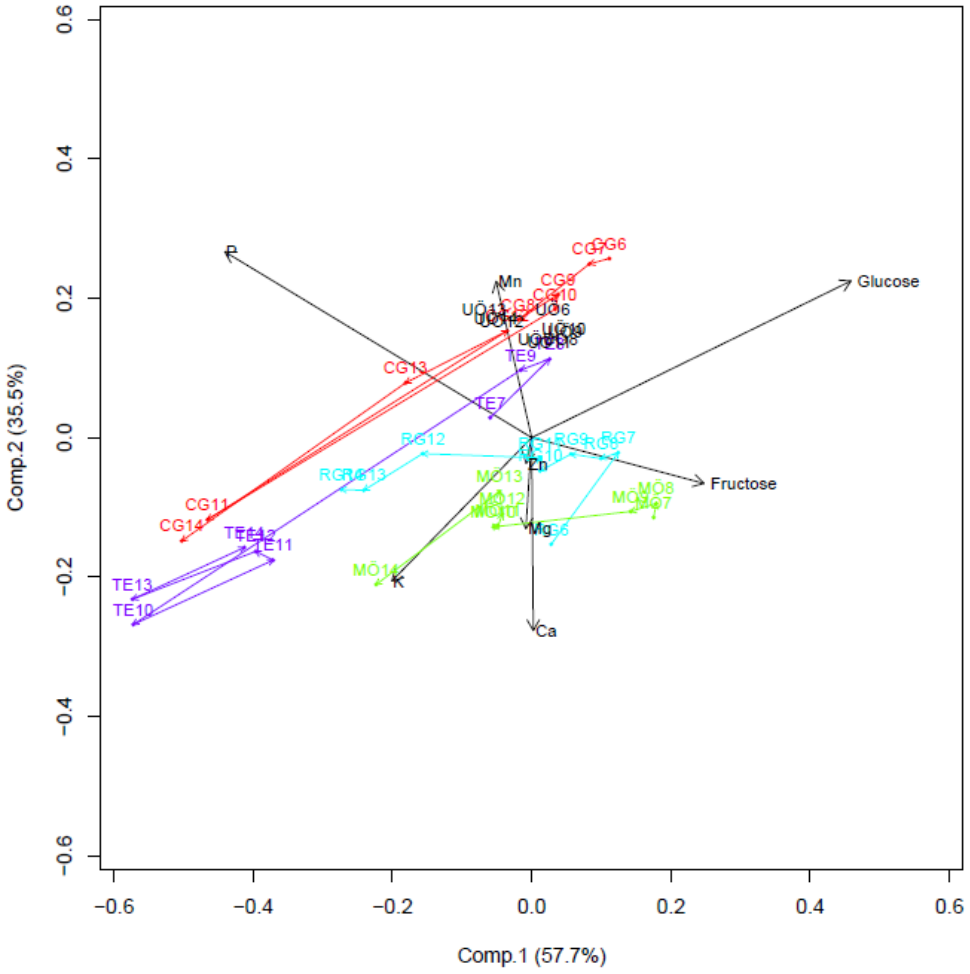


Figure 1: Biplot of sap season 2016. Five different series (trees) are connected with arrows showing the time evolvement. Samples are identified with initials from the forest owner (MÖ, TE, CG, UÖ and RG) combined with the day-number (6-14).

Figure 1 show decreasing trends in sugars, especially glucose, due to the movement of the time-series towards the left-bottom corner. Series UÖ is showing constant sap quality over the whole sampling period (no movement), while the CG-series fluctuates between high and low levels of fructose.

The biplot from year 2017 is shown in Figure 2.

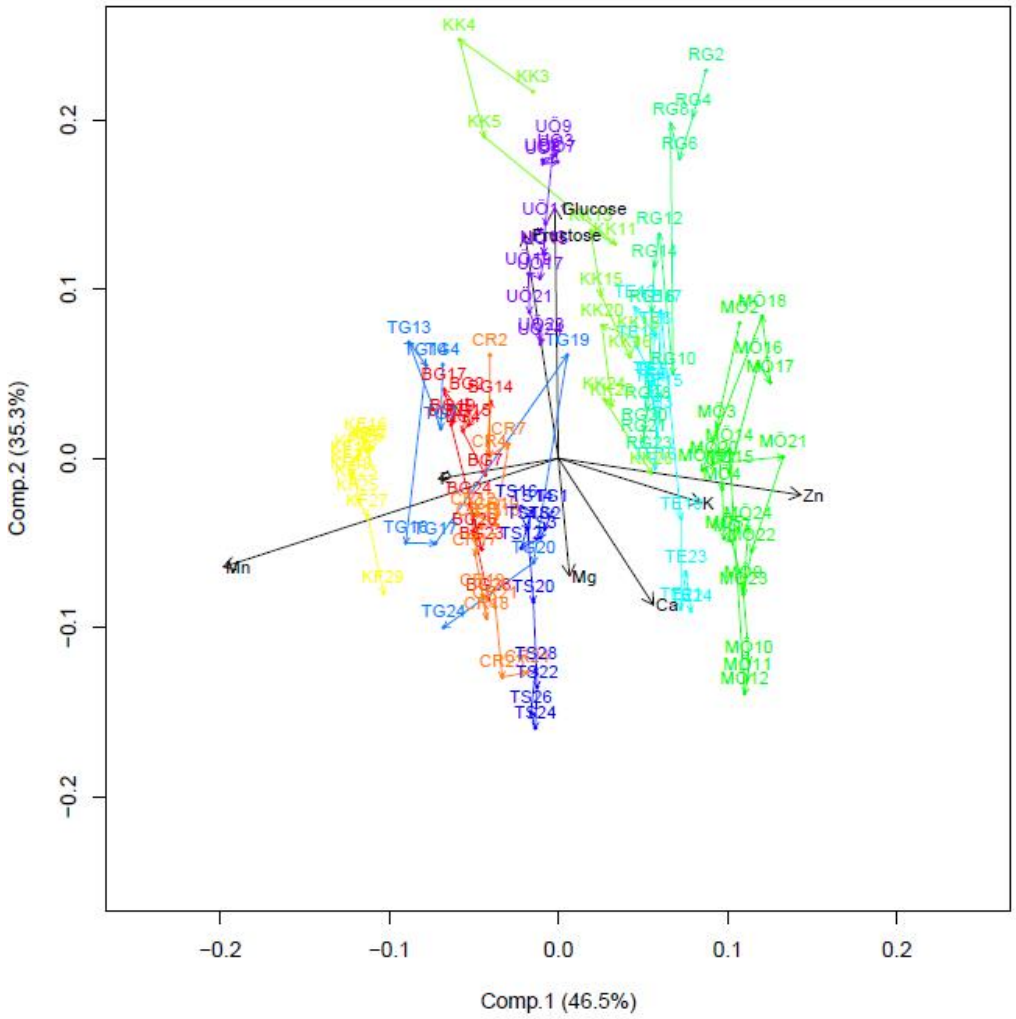


Figure 2: Biplot of sap season 2017. Ten different series (10 trees, 129 samples) are connected with arrows showing the time evolvement. Samples are identified with initials from the forest owner (TE, RG, MÖ, UÖ, KF, KK, CR, TS, BG and TG) combined with the day-number (1-29).

Figure 2 show strong correlation between glucose and fructose in the second (vertical) component. Again, the time-series evolve in the opposite direction to the glucose-fructose loading showing a decreasing trend in sugars. Metals Mg and Ca show an increasing trend towards the end of the sap season and Mn, P, K and Zn dominating the first component separating the series horizontally. MÖ having relatively high concentrations in Zn and K, while series KF having relatively high concentration of Mn and P. Having much more samples compared to 2016 gives a more reliable model compared to 2016.

Figure 3 show the result from sap season 2018. Series MÖ and KK showing high relative sugar concentration overs the whole sap period, while KF and CR being low in sugars. The same trend in decreasing sugars can also be seen here. Elements Mg and K are now more correlated to sugar content compared to previous years.

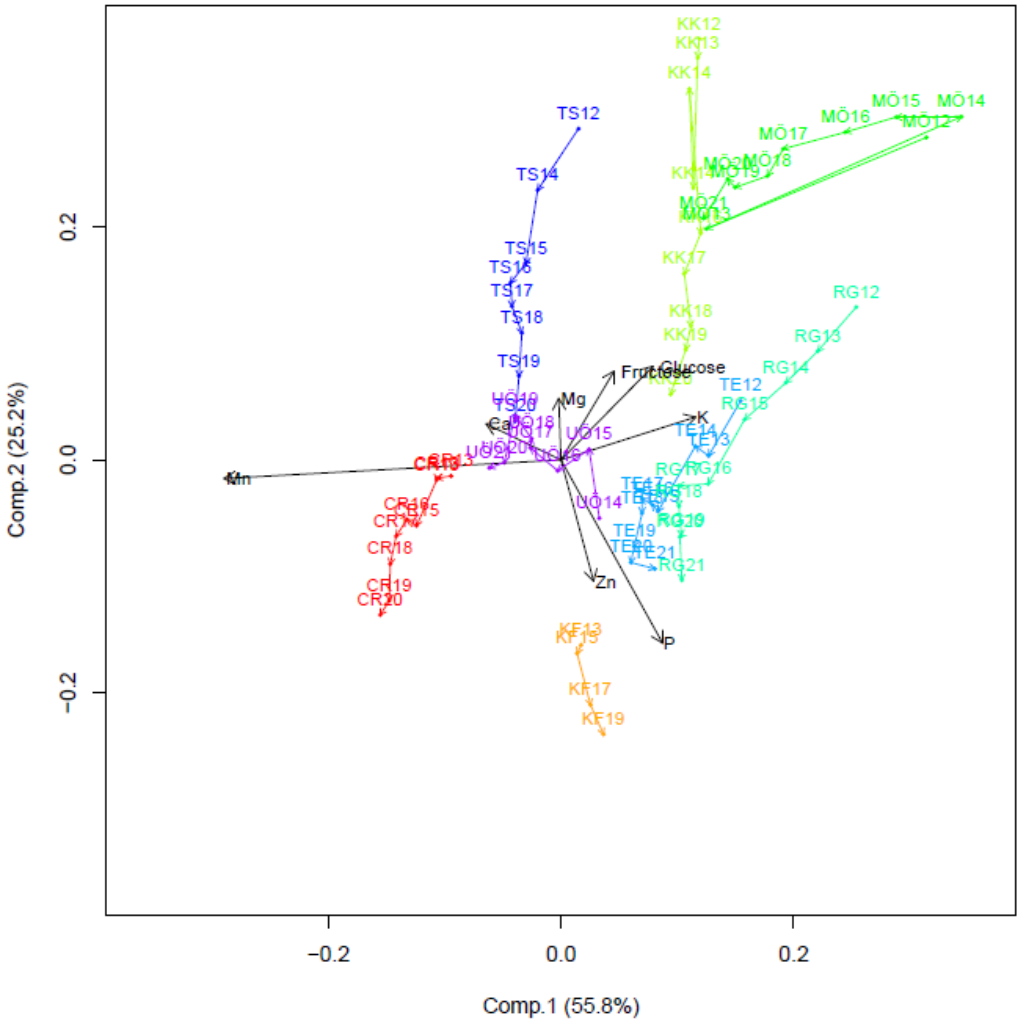


Figure 3: Biplot of sap season 2018. Eight different series (8 trees, 68 samples) are connected with arrows showing the time evolvement. Samples are identified with initials from the forest owner (TE, RG, MÖ, UÖ, KF, KK, CR and TS) combined with the day-number (12-21).

Comparing the results from all three years, we can see a general decreasing trend in relative sugar concentrations and an increasing trend in some of the metals. To get the big picture, a PCA model including all samples is calculated. In this model the parameters pH and conductivity is also included as external parameters. The influence of weather conditions is also of interest. Daily precipitation (mm) and daily average temperature (degC) was acquired from the closest available weather station which was in Vaasa Klemettilä (about 30 km away from the trees). It is believed that the amount of sunshine will affect

the production of birch sap. The growth place of the tree and potential surrounding shadowing trees will of course also affect the amount of light received by the tree. There might also be a lag, both in the temperature and in the precipitation, before they have an impact on the production of birch sap. Therefore lag-parameters were calculated. Precipitation values were added over a period of time, i.e. p4 indicates the sum of precipitation of the present day and four consecutive days before. Likewise, for temperature, i.e. t3 means the average temperature of the present day and three days before. Calculated precipitation and temperature parameters are p0…p7 and t0…t7. These are also included in the PCA-model, together with pH and conductivity as external parameters. The biplot of this PCA-model is shown in Figure 4.

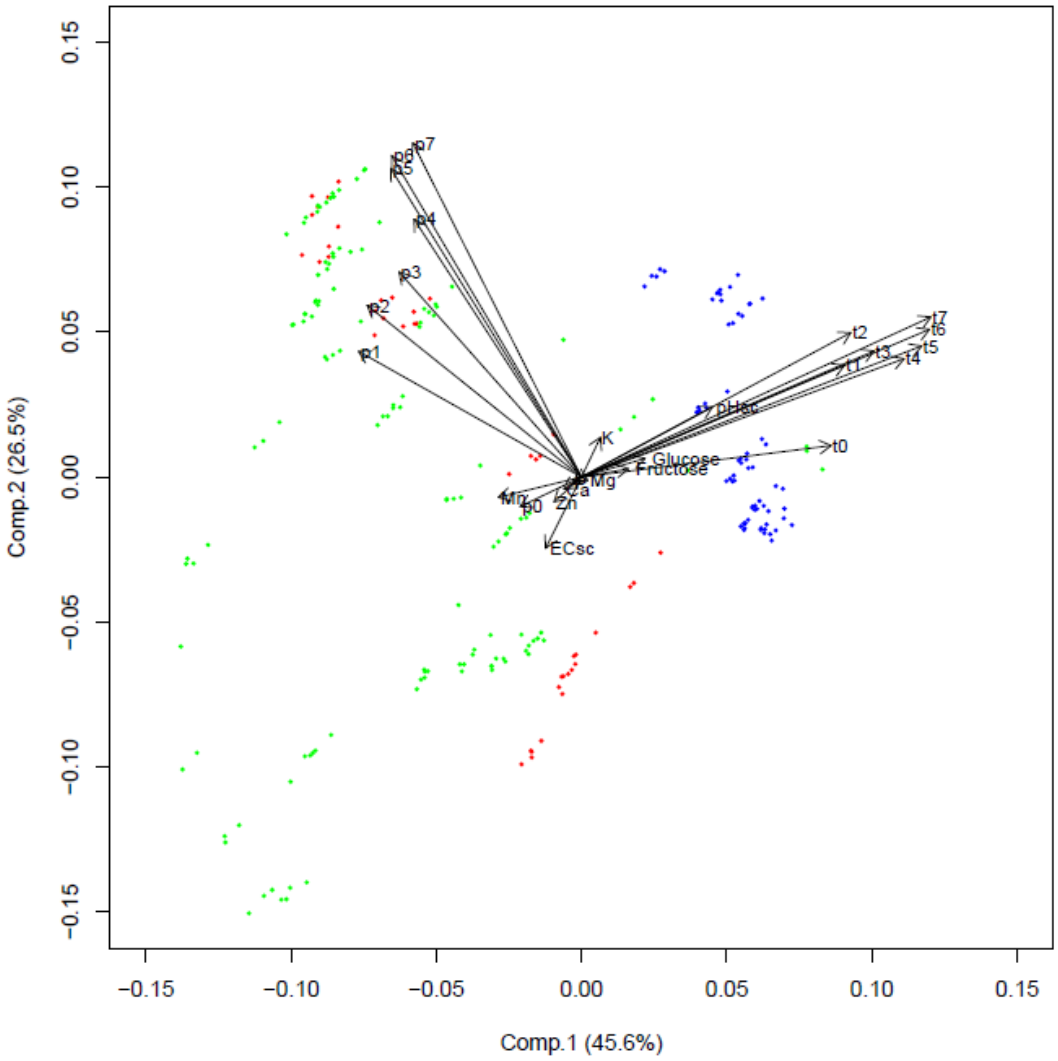


Figure 4: Biplot of sap season 2016-18, 2016 samples in red, 2017 in green and 2018 in blue.

Figure 4 shows interesting correlation between parameters. Fructose and glucose are strongly correlated with the temperature parameters. Hence, high temperatures during the sap period will increase the production of sugars. Higher temperatures usually also means more sunshine. Precipitation p0 shows a weak negative correlation with the sugars, suggesting that if it rains the excess water will slightly dilute the sugar concentrations in the sap. The lagged parameters p1-p7 are uncorrelated with sugars indicating

no long-term influence from precipitation. During this time of spring the soil is also quite humid because of the melted snow cover, thus small amounts of precipitation will not affect the sap process. In general, sap samples from year 2018 (blue in Figure 4) had higher relative concentrations of sugars compared to the other years. Spring 2018 was also the warmest period of all the three years. Most of the metals elements (Mn, Zn, Ca) are negatively correlated with pH, while P and Mg have low impact on the model. K is the only element that shows a positive correlation with the sugars. The big picture is that sugars, pH and to some extent element K are decreasing during the sap period, while the metals in general are increasing during the sap period. The decrease of pH is a result of increasing organic acids in the sap which was shown by Kallio (1987). Unfortunately, we were not able to measure the total daily production of birch sap per tree, but it can be assumed that in the beginning of the sap period the volumes are quite moderate, thus producing sugar rich concentrated birch sap, which is indicated by our results. If the goal is to collect "premium", sugar concentrated sap, it should be done in the beginning of the sap season.

Acknowledgements

The authors want to thank researcher Leif Hed at Centria Univeristy of Applied Sciences in Kokkola for doing all the chemical analyzes of the birch sap samples. This study was funded as a part of the European Regional Development Fund, Botnia-Alantica within the projects Industry Nordic and Bothnia Business Innovation.

References

Seal, R. (2016). Birch water: big business? Financial Times: https://www.ft.com/content/8590acdc-2dc8-11e6-a18d-a96ab29e3c95 (read 18.12.2018)

Kallio, H. and Ahtonen S. (1987). Seasonal variations of the sugars in birch sap. *Food Chemistry 25*, pp. 293-304.

Kallio, H. and Ahtonen S. (1987). Seasonal variations of the acids in birch sap. *Food Chemistry 25*, pp. 285-292.

Ahtonen S. and Kallio, H. (1989). Identification and seasonal variations of amino acids in birch sap used for syrup production. *Food Chemistry 33*, pp. 125-132.

Huldén S-G. and Harju L. (1986). Chemical analysis of mineral elements in spring sap of Birch. Daily and seasonal variations in the sap composition. *Acta Academiae Aboensis 46*(8).

R Core Team (2018) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Temple M., Hron K. and Filzmoser P. (2011). *robCompositions: an R-package for robust statistical analysis of compositional data*. In V. Pawlowsky-Glahn and A. Buccianti, editors, Compositional Data Analysis. Theory and Applications, pp. 341-355, John Wiley & Sons, Chichester (UK).

Biplots for Compositional Data Derived from Generalised Joint Diagonalisation Methods

U. Mueller[1], R. Tolosana Delgado[3], E.C. Grunsky[3] and J.M. McKinley[4]

[1]Center of Ecosystems Management, School of Science, Edith Cowan University Australia; u.mueller@ecu.edu.au

[2]Helmholtz Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resources Technology, Freiberg, Germany

[3]Waterloo University, Waterloo, Canada

[4]Queens University, Belfast, UK

Summary

Biplots constructed from principal components of a compositional data set are an established means to explore its features. Principal Component Analysis (PCA) is also used to transform a set of spatial variables into spatially decorrelated factors. However, because no spatial structures are accounted for in the transformation the application of PCA is limited. In Geostatistics and Blind Source Separation a variety of different matrix diagonalisation methods have been developed with the aim to provide spatially or temporally decorrelated factors. Just as PCA, many of these transformations are linear and so lend themselves to the construction of biplots. In this contribution we consider such biplots for a number of methods (MAF, UWEDGE and RJD transformations) and discuss how and if they can contribute to our understanding of relationships between the components of regionalised compositions. A comparison of the biplots with the PCA biplot commonly used in compositional data analysis for the case of data from the Northern Irish geochemical survey shows that the biplots from MAF and UWEDGE are comparable while that from RJD does not reveal any associations indicating that RJD might not be suitable for exploratory statistical analysis and that MAF might suffice to provide an adequate spatial characterisation.

Key words: semivariogram matrices, spatial decorrelation, structural analysis.

1 Introduction

Biplots constructed from principal components are an established means for exploring the features of a compositional data set. In several contributions (for example McKinley et al, 2018) we have seen that the method of minimum maximum autocorrelation factors (MAF, Switzer and Green, 1984) enhances classification and improves spatial decorrelation of factors derived from regionalised compositions. It is therefore often preferred in cases where

ISBN 978-84-947240-2-2

the regionalisation is important. However, to date biplots for MAF derived factors have not been explored, nor have biplots associated with more general joint diagonalisers based on the covariance or semivariogram function of the regionalised composition. These include Uniformly Weighted Exhaustive Diagonalisation with Gauss iterations (UWEDGE, Tichavsky and Yeredor, 2009) and Rotational Joint Diagonalisation (RJD, Cardoso and Souloumiac, 1996).

2 Joint diagonalisation

Without loss of generality we will assume that we are given a regionalised composition in some study area $\mathcal{A}$, $\{\boldsymbol{z}(u_\alpha) = [z_1(u_\alpha), \dots, z_K(u_\alpha)]: z_k(u_\alpha) > 0, k = 1, \dots, K; \sum_{k=1}^{K} z_k(u_\alpha) = 100, u_\alpha \in \mathcal{A}, \alpha = 1, \dots, n\}$, where $u_\alpha \in \mathcal{A}, \alpha = 1, \dots, n$ denotes a sample location. The corresponding clr transformed variables are $\{\boldsymbol{\zeta}(u_\alpha) = [\zeta_1(u_\alpha), \dots, \zeta_K(u_\alpha)]: k = 1, \dots, K, u_\alpha \in \mathcal{A},\ \alpha = 1, \dots, n\}$ with variance covariance matrix Σ_{clr} In addition the experimental semivariogram at lag h will be denoted by $\Gamma_{clr}(h) = cov\big(\boldsymbol{\zeta}(u), \boldsymbol{\zeta}(u+h)\big)\big)$, estimated from the sample locations. Thus associated with the regionalised composition there is a family $\{\Gamma_{clr}(h_\ell), \ell = 1, \dots, L\}$ of experimental semivariogram matrices calculated at L lag values h_ℓ chosen as appropriate to the nearest neighbor separation of the sample data. These describe the spatial continuity of the regionalised composition.

Since the covariance of the clr data is singular by construction, so are the semivariogram matrices, as a consequence working with ilr transformed variables is preferred. If V denotes the transformation from clr to ilr space, then the corresponding experimental semivariograms are given by $\Gamma_{ilr}(h_\ell) = V^T \Gamma_{clr}(h_\ell) V, \ell = 1, \dots, L$.

The general problem to be addressed is the following: Given a family of semivariogram matrices $M_\ell = \Gamma_{ilr}(h_\ell), \ell = 1, \dots, L$ find a matrix A such that for all ℓ the equation $M_\ell = A \Lambda_\ell A^T$ is valid where Λ_ℓ is a diagonal matrix. If such a matrix A exists, then the family $\{M_\ell: \ell = 1, \dots, L\}$ is said to be jointly diagonalizable.

The matrices are real symmetric by construction and there are two types of joint diagonalization that can be considered. One is joint diagonalisation via a similarity transformation, that is, the matrix A is orthogonal, or via a congruence, in which case the matrix A is no longer required to be orthogonal. The two types of diagonalization are termed orthogonal joint diagonalization (OJD) and non-orthogonal joint diagonalization (NOJD) respectively. Unless certain commutativity conditions on the matrices are satisfied, these techniques are approximate, when the number of matrices exceeds 1 for OJD and 2 for NOJD.

Given the regionalised composition the biplot for any one of the methods is constructed based on the clr-transformed data although the diagonalization matrix is derived from ilr variograms, indeed, if $\boldsymbol{Z}_{clr}(u)$ denotes the $n \times D$ matrix of centered clr-scores, then following Filzmoser et al (2009) the factors are given by

$$F(u) = Z_{clr}(u) V W$$

and the first two columns of $F(u)$ represent the scores and the first two rows of VW the loadings with W denoting the matrix derived by the diagonalization method.

2.1 PCA and MAF

The solution proposed by the PCA method consists of determining the eigenvalue

decomposition of the variance-covariance matrix M of the ilr transformed data, $M = W\Lambda W^T$. The matrices W and Λ are derived from an eigenvector eigenvalue decomposition and the eigenvalues in the matrix Λ are arranged in descending order. They reflect the variability represented by the corresponding factor.

For MAF, the variance-covariance matrix M and the semivariogram matrix $\Gamma_{ilr}(h)$ at a chosen lag h are diagonalised jointly by congruence: $AMA^T = I$ and $A\Gamma_{ilr}(h)A^T = \Lambda_1$. The matrix A is non-singular and is given by $A = W_1^T\Lambda^{-1/2}W$A where W and Λ are the orthogonal matrix and diagonal matrix derived from the eigenvalue decomposition of M and W_1 is the orthogonal matrix which diagonalises $\Lambda^{-1/2}W^T\Gamma_Z(h)W\Lambda^{-1/2}$. The eigenvalues in the matrices Λ and Λ_1 are arranged in descending order.

2.3 Joint approximate diagonalization

Diagonalisation via PCA is an OJD method, while MAF is a NOJD method. In the context of a family of semivariogram matrices that is sought to be diagonalised, the transformation derived from PCA will only diagonalise all semivariogram matrices if they commute pairwise. This condition is typically not satisfied, as the data on which calculation of the semivariogram matrices is based are noisy and there are different spatial scales that impact on the continuity of the data.

To account for phenomena of this type, Blind Source Separation attempts to derive the best diagonaliser based on the entire family of matrices available. Typically some kind of fixed point iteration is used to determine the matrix A that best jointly diagonalises the given family of symmetric matrices according to some cost criterion. In the case of OJD the cost function is set to be

$$C_1(A) = \sum_{\ell=1}^{L} trace((A^T\mathrm{M}_\ell A - diag(A^T\mathrm{M}_\ell A)^T\,(A^T\mathrm{M}_\ell A - diag(A^T\mathrm{M}_\ell A))$$

where one seeks to minimize the sum of squares of the off-diagonal entries in $A^T\Lambda_\ell A$. In the RJD algorithm the matrix A is constructed iteratively from Jacobi rotation matrices.

The criterion used for NOJD is not very different from it, in the case of the UWEDGE method a matrix A is sought that minimises

$$C_3(A,W) = \sum_{\ell=1}^{L} trace((W^T\mathrm{M}_\ell W - A\Lambda_{\ell,V}A^T)^T\,(W^T\mathrm{M}_\ell W - diag\left(A\Lambda_{\ell,V}A^T\right))$$

Here $\Lambda_{\ell,W} = \mathrm{diag}(W^T\mathrm{M}_\ell W)$) and the matrices A and W are called the mixing matrix and demixing matrix respectively. The resulting loading matrices are summarized in Table 1.

Table 1: Loading matrices for diagonalisation methods

Method	clr-loading matrix
PCA	VW
MAF	$VW\Lambda^{-1/2}W_1$
RJD	VA
UWEDGE	VW^T

ISBN 978-84-947240-2-2

3 Data

The Northern Irish Tellus Survey (GSNI, 2007; Young and Donald, 2013) consists of 6862 rural soil samples (X-ray fluorescence (XRF) analyses). Geochemical samples presented in this study were collected at 20-cm depth, with average spatial coverage of one sample site every 2 km^2. Each soil sample site was assigned one of six broad lithological classes (acid volcanics, felsic magmatics, basic volcanics, mafic magmatics, carbonatic, silicic clastics) and at each location 50 continuous geochemical variables were retained for analysis (Ag, Al_2O_3, As, Ba, Bi, Br, CaO, Cd, Ce, Cl, Co, Cr, Cs, Cu, Fe_2O_3, Ga, Ge, Hf, I, K_2O, La, LOI, MgO, MnO, Mo, Na2O, Nb, Nd, Ni, P_2O_5, Pb, Rb, SO_3, Sb, Sc, Se, SiO_2, Sm, Sn, Sr, Th, TiO_2, Tl, U, V, W, Y, Yb, Zn, Zr and Loss on Ignition (LOI). More information on Tellus Survey field methods and analytical methodology are available in Smyth (2007) and Young and Donald (2013).

To illustrate the methods, biplots constructed from 2 subcompositions were considered. The first is the subcomposition comprised of the oxides Al_2O_3, CaO, Fe_2O_3, K_2O, MgO, MnO, Na2O, P_2O_5, SiO_2, TiO_2, and LOI and the second is that consisting of oxides and selected trace elements as described in Tolosana-Delgado and McKinley (2016).

For each of the two sets a default ilr transform was performed and experimental direct and cross variograms were computed for 30 lags at a nominal spacing of 1 km. The MAF transform was based on an estimate of the covariance matrix and the semivariogram matrix for the first lag, for the RJD and UWEDGE methods the semivariogram matrices for all lags up to distance 20 km were used.

4 Results

The experimental semivariograms of the factors and biplots of the first two components for the subcomposition of major oxides are shown in Figure 1. The experimental semivariograms for the two methods show strong similarities with only minor differences between the variograms of the same index. For PCA and RJD no ordering by spatial continuity results as a consequence of the transformation. For PCA, the experimental semivariogram of the second factor is similar in shape to the one of the first factor of MAF or UWEDGE. The loadings of these factors (maf1, uwedge1, pca2) show a similar interpretation as a broad balance between mafic elements and felsic elements, which is related to the contrast of the Antrim basalts against virtually the rest of Northern Ireland, as can be seen in maps of the corresponding scores (not shown for brevity). For higher order factors of MAF and UWEDGE, variograms show a progressive destructuralization, with decreasing range and increasing nugget to sill ratio. This is reflected in their score maps, where decreasing scales of high and low value regions and increasing noise are evident. Neither PCA nor RJD share this behavior. In fact, all but the PCA variograms show the same structure, while RJD variograms have two basic shapes, one associated with the large scale continuity, the other with shorter scale, and no natural ordering.

Biplots for MAF and UWEDGE are almost identical, showing a ternary system of mafics, felsics, and silicic clastic materials. So UWEDGE is not necessary here. Comp1 for PCA is related to peat building (Tolosana-Delgado and McKinley, 2016). All three show a separation of lithologies, which is absent in RJD biplot. This may be related to the lack of natural order in the RJD structures, making RJD unsuitable for exploratory analysis.

ISBN 978-84-947240-2-2

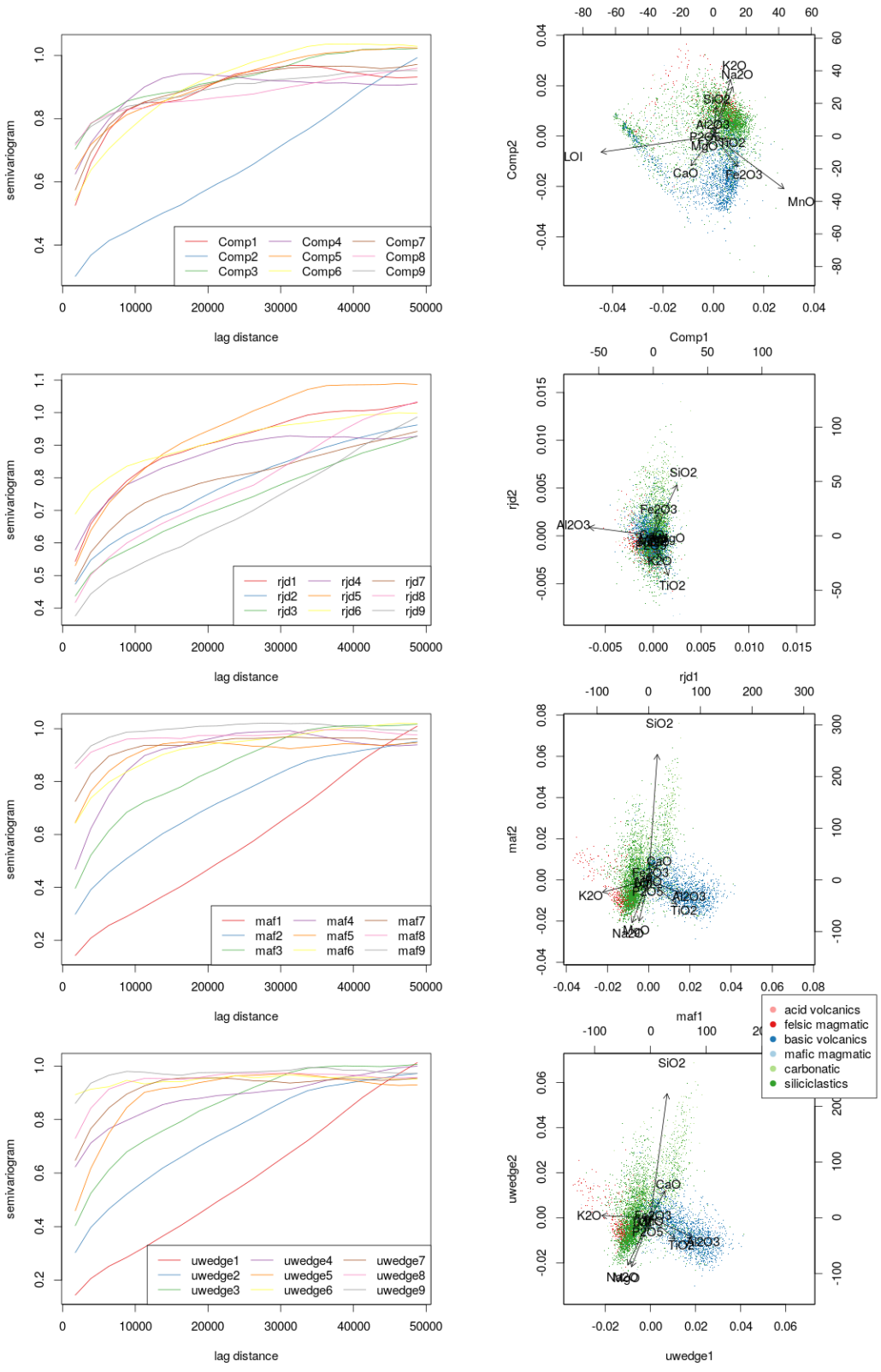


Figure 1: Experimental semivariograms of factors (left) and biplots coloured by lithology (right) by method (order from top to bottom: PCA, RJD, MAF, UWEDGE)

ISBN 978-84-947240-2-2

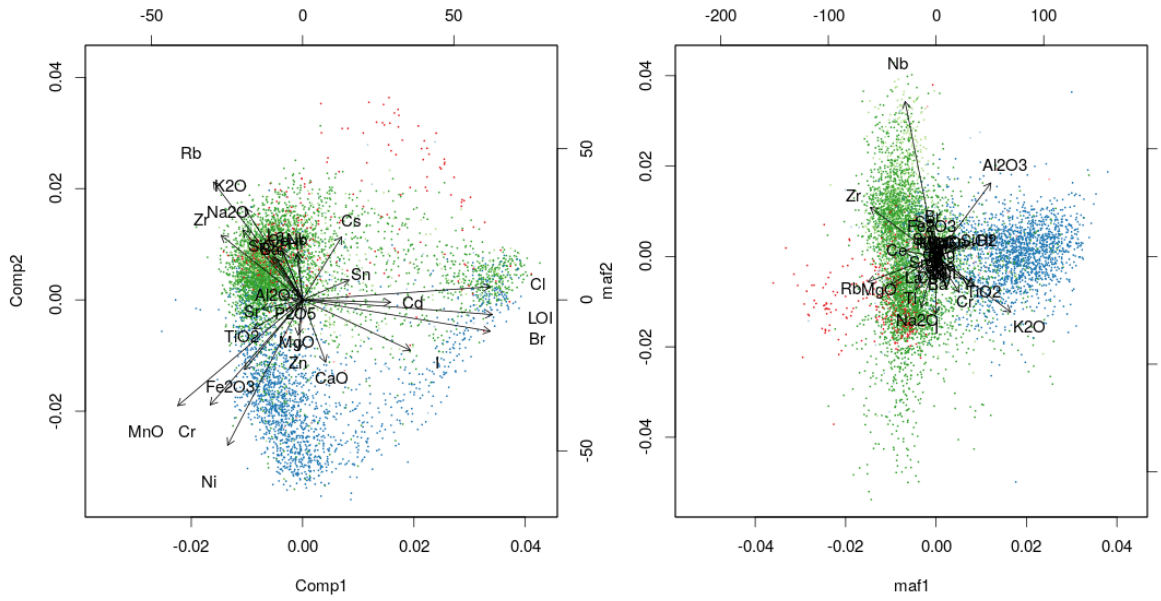


Figure 2: Biplots coloured by lithology for the extended set of variables by method: PCA (left) and MAF (right)

The second subcomposition shows a similar structure in the PCA biplot, while in the MAF biplot the scores do not change very much but the loadings are dominated by different variables, mostly trace elements. Variograms show similar behavior as for the first subcomposition.

References

Cardoso, J. K. and Souloumiac A. (1996). Jacobi angles for simultaneous diagonalisation. *SIAM Journal of Matrix Analysis and Applications* 17, pp. 161-164.

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

GSNI, (2007). Geological Survey Northern Ireland Tellus project overview. https://www.bgs.ac.uk/gsni/Tellus/index.html. Accessed 7 Mar 2017.

McKinley, J., Grunsky E. and Mueller U. (2018). Environmental Monitoring and Peat Assessment Using Multivariate Analysis of Regional Scale Geochemical Data, *Mathematical Geosciences* 50, pp. 235–246.

Smyth, D, (2007) Methods used in the Tellus geochemical mapping of Northern Ireland. British Geological Survey Open Report or/07/022.

Switzer P. and Green A.A. (1984). Min/Max autocorrelation factors for multivariate spatial imaging, Palo Alto, California: Stanford University.

Tichavsky, P. and Yeredor, A. (2009). Fast approximate joint diagonalization incorporating weight matrices. *IEEE Trans on Sig Proc* 57, pp. 878 – 891.

Tolosana-Delgado, R. and McKinley J. (2016). Exploring the joint compositional variability of major components and trace elements in the Tellus soil geochemistry survey (Northern Ireland), *Applied Geochemistry* 75, pp. 263-276.

Young, M and Donald, A, 2013, A guide to the Tellus data. Geological Survey of Northern Ireland, Belfast.

ISBN 978-84-947240-2-2

CODA methods and the multivariate Student distribution: an application to political economy

T.H.A. Nguyen[1], T. Laurent[2]
[1]Toulouse School of Economics, France; *huongan.nguyen@tse-fr.eu*
[2]Danang University of Architecture, Vietnam

Summary

In a multiparty election, the vote shares form a composition vector (mathematically, a vector belonging to a simplex). Political economists are interested by the impact of the characteristics of the geographical units on the outcome of the elections. Because vote shares data often exhibit heavy tail behavior, we decide to use a Student error distribution. We describe how to adapt the CODA regression model to the multivariate Student error distribution. For a Gaussian errors vector, the assumption of independent coordinates is equivalent to the assumption of correlated coordinates. However, this equivalence is no longer true when considering a multivariate Student distribution. In this paper, we recall these two types of multivariate Student distribution for the error term, and concentrate on building a CODA regression model using the multivariate independent Student error vectors. We compare this model to a model which uses the multivariate Gaussian distribution. The models are fitted on French electoral data of the 2015 departmental elections. We illustrate on this data set a method for selecting between the Gaussian and the Student models based on the Mahalanobis distance.

Key words: Independent multivariate Student distribution, Uncorrelated multivariate Student distribution, compositional regression models, Maximum Likelihood Estimator, heavy tail, R ..

1 Introduction

Recently, a lot of authors in policital economy concentrate on building models and understanding the drivers of the outcome of a two-party electoral system (Beauguitte and Colange (2013), Ansolabehere and Leblanc (2008)). The outcome of an election can be influenced by the campaign strategies of candidates, demographic factors such as age, domain of activity, rate of unemployment, and so on. In an interview with Time magazine, a group of Obama senior campaign advisers revealed an enormous data effort to support fundraising, micro-targeting TV ads and modeling of swing-state voters. In this work, we are interested in exploring the impact of the characteristics of the demographics and social factors on the outcome of the 2015 French departmental election. The outcomes of the election in this multiparty system consist of vectors whose components are the percentages of proportions of votes per party. In what follow, my attention focuses on the relation between votes shares and socio-economics factors such as age, education levels, domain of activities, unemployment rate and so on by using a CODA (COmpositional Data Analysis) regression models.

Among papers concentrating on the relationship between socio-economic variables and election outcomes, Beauguitte and Colange (2013) carry out a linear regression model at three levels of aggregation (polling stations, cities, and electoral districts) and show that the socio-economic variables are significant. Kavanagh et al. (2006) use a geographically weighted regression, which produces parameters estimates for each data

ISBN 978-84-947240-2-2

point, i.e. for each electoral division. In the statistical literature, there are regression models adapted to share vectors including CODA models, but also Dirichlet models, Student model and others. These models, where the dependent and independent variables may be compositional variables (see Mert et al. (2018)). Honaker et al. (2002), Katz and King (1999) use a statistical model for multiparty electoral data assuming that the territorial units yield independent observations. Morais et al. (2017) studies the impact of media investments on brand's market shares with a CODA regression model. Nguyen et al. (2018) study a CODA multivariate regression model which uses the normal distribution to illustrate the impacts of socio-economic factor on French departmental election. However, this election data often exhibit heavy tail behavior (see Katz and King (1999)). In order to eliminate the heavy tail problem, a proposal found in the literature is to replace the Gaussian distribution by the Student distribution in this paper.

In one dimension, the generalized Student distribution encompasses the gaussian distribution as a limit when the degrees of freedom tends to infinity, allowing for heavier tails when the shape parameter is small. However in higher dimensions, there are several kinds of multivariate Student models (see Johnson and Kotz (1972) and Kotz and Nadarajah (2004) for overview). There are two versions of Student distribution: the independent Student (IT) and the uncorrelated Student (UT) (see Kelejian and Prucha (1985)). There are some authors who concentrate on the univariate Student while others pay attention in multivariate Student. Nguyen et al. (2019) perform a full summary of these versions and consider a multivariate dependent vector and a linear regression model with three different assumptions on the error term distribution: the Gaussian distribution ($\boldsymbol{\epsilon}_N$), the Uncorrelated Student distribution ($\boldsymbol{\epsilon}_{UT}$), the Independent Student distribution ($\boldsymbol{\epsilon}_{IT}$). Nguyen et al (2019) derive some theoretical properties of the UT model and propose a simple iterative reweighted algorithm to compute the maximum likelihood estimators in the IT model. However, Nguyen et al. (2019) show that the UT model is simpler to fit than the IT model, but it has limitation of assumption of the single realization. This restricts the properties of the maximum likelihood estimators and prevent the use of tests against the other two models. Thus, we will concentrate in multivariate IT case and compare it to multivariate Gaussian case in this paper.

Section 2 decribes the departmental election data. Section 3 presents the multivariate regression models (includes multivariate Normal error vector and multivariate Independent Student (IT) error vector). In section 4, we recall the CODA principles then build a CODA regression model that could be considered to explain the outcome of an election and to clarify its relations with the socio-economic factors by using the independent multivariate Student distribution (IT) with known degree of freedom. As in Nguyen et al. (2019), we perform a test based on the Mahalanobis distance to select between the multivariate Gaussian and the multivariate Student models in Section 5.

2 Data

Vote share data of the 2015 French departmental election for 95 departments in France are collected from the Cartelec website [1] and corresponding socio-economic data (for 2014) have been downloaded from the INSEE website [2]. Table 1 summarizes our data set.

Employment has five categories: AZ (agriculture, fisheries), BE (manufacturing industry, mining industry and others), FZ (construction), GU (business, transport and services) and OQ (public administration, teaching, human health). Diploma has three levels: <BAC for people with at most some secondary education, BAC for people with at least some secondary education and at most a high school diploma, and SUP for people with a university diploma. The Age variable has three levels: Age_1840 for people from 18 to 40 years old, Age_4064 for people from 40 to 64 years old, and Age_65 for elderly. For the vote share variable, the

[1] https://www.data.gouv.fr/fr/datasets/elections-departementales-2015-resultats-par-bureaux-de-vote/
[2] https://www.insee.fr/fr/statistiques

Table 1: Data description.

Variable name	Description	Averages
Vote share	Left(L), Right(R), Extreme Right(XR)	0.37, 0.388, 0.242
Age	Age_1840, Age_4064, Age_65.	0.313, 0.432, 0.255
Diploma	<BAC, BAC, SUP.	0.591, 0.16, 0.239
Employment	AZ, BE, FZ, GU, OQ	0.031, 0.099, 0.049, 0.439, 0.382
unemp	The unemployment rate	0.117
employ_evol	Mean annual growth rate of employment (2009-2014)	-0.145
owner	The proportion of people who own assets	0.616
income_ tax	The proportion of people who pay income tax	0.552
foreign	The proportion of foreigners	0.050

Cartelec website provides a very detailed information. The list of political parties which present candidates at that election is higher than 15. In France, the Ministry of the Interior is in charge of publishing the electoral results. Despite the dissatisfaction of some political parties, the Ministry of the Interior summarized the results by grouping the political parties into three main components : Left, Right and Extreme-Right[3]. Note that the averages in the last column of Table 1 are geometric means by component in the 324 case of compositional variables.

From the CODA point of view, when compositional data have three components, they can be represented in a ternary diagram. For instance, the vote shares of the 95 departments for the Left and Right wings and the Extreme Right party are the black points in Figure 4. The red triangle corresponding to the Aube department on Figure 4 shows that its vote shares of the Left wing, the Right wing and the Extreme Right party are respectively 17.4%, 54.6%, and 28% . Figure 2 illustrates the positions of the French departments on the ternary diagram whose components correspond to the three levels of the diploma variable, and the red triangle figures the geometric mean (adapted mean for compositional data) of all departments.

3 The multivariate regression models

Let us consider a model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \tag{1}$$

where $\mathbf{Y}$ is a $n \times L$ matrix of L dimensional dependent variable, $\mathbf{X}$ is a $n \times (K+1)$ matrix of K explanatory variables, $\boldsymbol{\beta}$ is the parameter matrix of size $(K+1) \times L$ and $\boldsymbol{\epsilon}$ is the error matrix of size $n \times L$.

3.1 Multivariate Normal error vector

Let us first consider model (**??**) with independent and identically distributed error vectors $\boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$, following a multivariate normal distribution $\mathcal{N}_L(\mathbf{0}, \boldsymbol{\Sigma})$ with an L-vector of means equal to zero and an $L \times L$

[3]for more details, see https://fr.wikipedia.org/wiki/Elections_départementales_francaises_de_2015

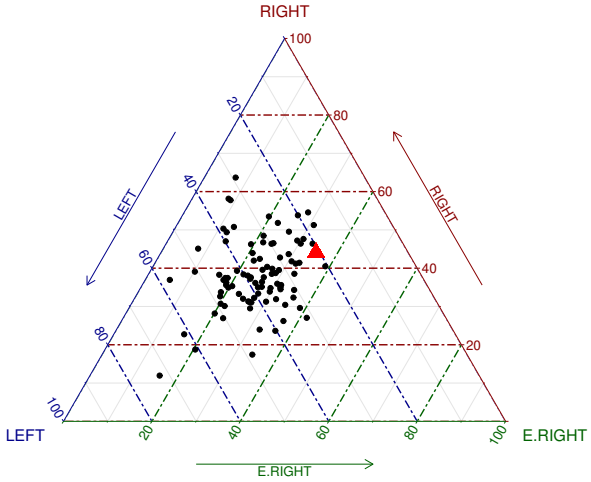


Figure 1: Vote shares in the 95 departments (blue points) with the Aube department (red triangle).

covariance matrix $\boldsymbol{\Sigma}$. This model is denoted by N and the subscript N is used to denote the error terms $\boldsymbol{\epsilon}_{Ni}$, $i = 1, \ldots, n$ and the parameters $\boldsymbol{\beta}_N$ and $\boldsymbol{\Sigma}_N$ of the model. The maximum likelihood estimators of $\boldsymbol{\beta}_N$ and $\boldsymbol{\Sigma}_N$ are

$$\hat{\boldsymbol{\beta}}_N = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{2}$$

$$\hat{\boldsymbol{\Sigma}}_N = \frac{\sum_{i=1}^{n} \hat{\boldsymbol{\epsilon}}_{Ni}\hat{\boldsymbol{\epsilon}}_{Ni}^T}{n}, \tag{3}$$

where $\hat{\boldsymbol{\epsilon}}_{Ni} = \mathbf{Y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}_N$ (see e.g. Theorem 8.4 from Seber (2009)).

The estimator $\hat{\boldsymbol{\beta}}_N$ is an unbiased estimator of $\boldsymbol{\beta}_N$ while the bias of $\hat{\boldsymbol{\Sigma}}_N$ is equal to $-((K+1)/n)\boldsymbol{\Sigma}_N$ and tends to zero when n tends to infinity (see e.g. Theorems 8.1 and 8.2 from Seber (2009)).

3.2 Multivariate Independent Student error vector

According to Nguyen et al. (2019), let us denote the L dimensional dependent vector by:

$$\boldsymbol{\mathcal{Y}}_i = (y_{i1}, \ldots, y_{iL})^T.$$

For K explanatory variables, the design matrix is of size $L \times (K+1)L$ and is given by:

$$\boldsymbol{\mathcal{X}}_i = \boldsymbol{I}_L \otimes \boldsymbol{x}_i^T$$

for $i = 1, \ldots, n$, with the $(K+1)$-vector $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{iK})^T$, $\boldsymbol{I}_L$ the identity matrix with dimension L and $\otimes$ the usual Kronecker product. The parameter of interest is a $(K+1)L$ vector given by:

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_L^T)^T,$$

where $\boldsymbol{\beta}_j = (\beta_{0j}, \ldots, \beta_{Kj})^T$, for $j = 1, \ldots, L$ and the L-vector of errors is denoted by:

$$\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iL})^T$$

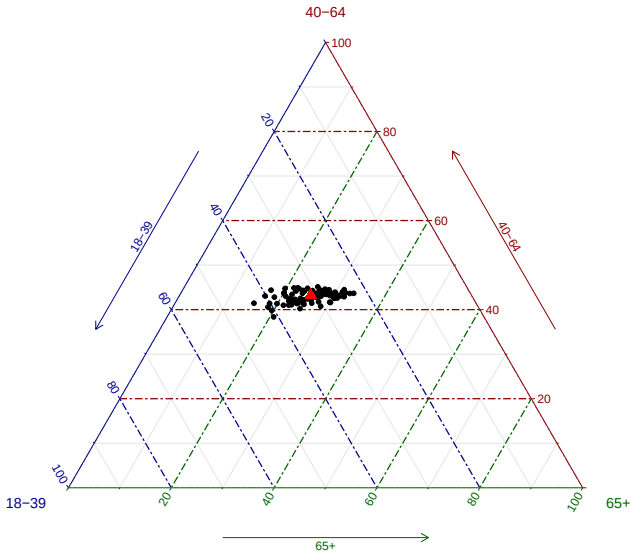


Figure 2: Ternary diagram of Age in the 95 departments

for $i = 1, \ldots, n$. Given these notations, model (**??**) can be rewritten as

$$\boldsymbol{\mathcal{Y}}_i = \boldsymbol{\mathcal{X}}_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i \tag{4}$$

with $\mathbb{E}(\boldsymbol{\epsilon}_i) = 0$ and $i = 1, \ldots, n$. Note that $\boldsymbol{\mathcal{Y}} = (\boldsymbol{\mathcal{Y}}_1^T, \ldots, \boldsymbol{\mathcal{Y}}_n^T)^T$ and $\boldsymbol{\epsilon} = (\boldsymbol{\epsilon}_1^T, \ldots, \boldsymbol{\epsilon}_n^T)^T$ are nL vectors, $\boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{X}}_1^T, \ldots, \boldsymbol{\mathcal{X}}_n^T)^T$ is the $nL \times (K+1)L$ matrix.

Let us consider model (4) with i.i.d. $\boldsymbol{\epsilon}_i$, $i = 1, \ldots, n$, following a independent multivariate Student (IT) distribution with L dimensions and known degrees of freedom $\nu > 2$. In most of the literature on multivariate Student, the density is rather parametrized as a function of the scatter matrix $((\nu-2)/\nu)\boldsymbol{\Sigma}$.

The probability density function for a L-vector $\boldsymbol{\epsilon}$

$$p(\boldsymbol{\epsilon}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{f(\nu)}{\det(\boldsymbol{\Sigma})^{1/2}} \left[1 + \frac{1}{\nu - 2}(\boldsymbol{\epsilon} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\epsilon} - \boldsymbol{\mu})\right]^{-(\nu+p)/2}, \tag{5}$$

where ^{T} denotes the transpose operator, $f(\nu) = \dfrac{\Gamma[(\nu+p)/2]}{\Gamma(\nu/2)(\nu-2)^{p/2}\pi^{p/2}}$ and Γ is the usual Gamma function. Following Prucha and Kelejian (1984), Nguyen et al. (2019) derive the maximum likelihood estimators for the IT model. The maximum likelihood estimators of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ in the IT regression model satisfy the following

implicit equations:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{IT} &= \left(\sum_{i=1}^{n} \hat{w}_{ITi} \boldsymbol{\mathcal{X}}_i^T \hat{\boldsymbol{\Sigma}}_{IT}^{-1} \boldsymbol{\mathcal{X}}_i\right)^{-1} \sum_{i=1}^{n} \hat{w}_{ITi} \boldsymbol{\mathcal{X}}_i^T \hat{\boldsymbol{\Sigma}}_{IT}^{-1} \boldsymbol{\mathcal{Y}}_i \\ \hat{\boldsymbol{\Sigma}}_{IT} &= \frac{1}{n} \sum_{i=1}^{n} \hat{w}_{ITi} \hat{\boldsymbol{\epsilon}}_{ITi} \hat{\boldsymbol{\epsilon}}_{ITi}^T \end{aligned} \tag{6}$$

$$\text{with} \quad \hat{\boldsymbol{\epsilon}}_{ITi} = \boldsymbol{\mathcal{Y}}_{\boldsymbol{i}} - \boldsymbol{\mathcal{X}}_i \hat{\boldsymbol{\beta}}_{IT} \quad \text{and} \quad \hat{w}_{ITi} = \frac{\nu + L}{\nu - 2 + \hat{\boldsymbol{\epsilon}}_{ITi}^T \hat{\boldsymbol{\Sigma}}_{IT}^{-1} \hat{\boldsymbol{\epsilon}}_{ITi}}.$$

We use the iterative reweighted algorithm as in Nguyen et al (2019) to estimate the coefficient and variance-covariance matrix.

4 Compositional regression models

4.1 Principles of compositional data analysis

4.1.1 Definition and operations

A composition $\mathbf{x}$ is a vector of D parts of some whole which carries relative information. A D-composition $\mathbf{x}$ lies in the so-called simplex space $\mathbf{S}^D$ defined by:

$$\mathbf{S}^D = \{\mathbf{x} = (x_1, ..., x_D)' : x_j > 0, j = 1, ..., D; \sum_{j=1}^{D} x_j = 1\}$$

Let $\mathcal{C}(\mathbf{x}) = \left(\frac{x_1}{\sum_{j=1}^{D} x_j}, \cdots, \frac{x_D}{\sum_{j=1}^{D} x_j}\right)$ is the closure operation, the vector space structure of the simplex $\mathbf{S}^D$ is defined by the perturbation and powering operations:

$$\begin{aligned}&\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D), \ \mathbf{x}, \ \mathbf{y} \in \mathbf{S}^D \\ &\lambda \odot \mathbf{x} = \mathcal{C}(x_1^{\lambda}, \ldots, x_D^{\lambda}), \ \lambda \text{ is a scalar}, \mathbf{x} \in \mathbf{S}^D.\end{aligned}$$

The compositional matrix product, corresponding to the matrix product in the simplex, is defined by

$$\mathbf{B} \boxdot \mathbf{x} = \mathcal{C}\left(\prod_{j=1}^{D} x_j^{b_{1j}}, \cdots, \prod_{j=1}^{D} x_j^{b_{Lj}}\right)^T$$

where $\mathbf{B} = (b_{lj}), \ l = 1, \ldots, L, \ j = 1, \ldots, D$, is a parameter matrix such that the column vectors belong to $\mathbf{S}^D, \mathbf{j}_L^T \mathbf{B} = \mathbf{0}_D, \mathbf{B}\mathbf{j}_D = \mathbf{0}_L$, where $\mathbf{j}_L$ is a $L \times 1$ column vector of ones, and $\mathbf{j}_L^T$ is the transposed of $\mathbf{j}_L$.
The simplex $\mathbf{S}^D$ can be equipped with the Aitchison inner product (see Aitchison (1982) and Pawlowsky et al (2015)) in order to define distances. The expected value $\mathbb{E}^{\oplus}\mathbf{Y}$ are also defined in Pawlowsky et al (2015).

4.1.2 Log-ratio transformation

Classical regression models cannot be used directly in the simplex because the constraints that the components are positive and sum up to 1 are not compatible with their usual distributional assumptions. To overcome this difficulty, one way out is to use a log-ratio transformation from the simplex space $\mathbf{S}^D$ to the Euclidean space $\mathbb{R}^{D-1}$. The classical transformations are alr (additive log-ratio transformation), clr (centered log-ratio transformation), and ilr (isometric log-ratio transformation). The coordinates in the clr transformed vector are linearly dependent, and the coordinates in the alr transformed vector are not compatible with the geometry (distance between the components in the simplex space is different from distance between the coordinates in the Euclidean space). For these reasons people generally use one of the ilr transformation for compositional regression models.
An isometric log-ratio transformation ilr is defined by:

$$\text{ilr}(\mathbf{x}) = \mathbf{V}_D^T \ln(\mathbf{x})$$

where the logarithm of $\mathbf{x}$ is understood componentwise, $\mathbf{V}_D^T$ is a transposed contrast matrix Pawlowsky et al (2015) associated to a given orthonormal basis $(\mathbf{e}_1, \cdots, \mathbf{e}_{D-1})$ of $\mathbf{S}^D$ by

$$\mathbf{V}_D = \text{clr}(\mathbf{e}_1, \cdots, \mathbf{e}_{D-1}).$$

As in Pawlowsky et al (2015) in our application, we use the following contrast matrix for $D = 3$

$$\mathbf{V}_3 = \begin{bmatrix} 2/\sqrt{6} & 0 \\ -1/\sqrt{6} & 1/\sqrt{2} \\ -1/\sqrt{6} & -1/\sqrt{2} \end{bmatrix}$$

This particular matrix defines the following ilr coordinates

$$\text{ilr}_1(\mathbf{x}) = \frac{1}{\sqrt{6}}(2\log x_1 - \ln x_2 - \log x_3) = \frac{2}{\sqrt{6}} \log \frac{x_1}{\sqrt{x_2 x_3}}$$
$$\text{ilr}_2(\mathbf{x}) = \frac{1}{\sqrt{2}}(\log x_2 - \log x_3) = \frac{1}{\sqrt{2}} \log \frac{x_2}{x_3}$$

The first ilr coordinate contains information about the relative importance of the first component x_1 with respect to the geometric mean of the second and the third components $g = \sqrt{x_2 x_3}$. The second ilr coordinate contains information about the relative importance of the second component x_2 with respect to the third component x_3. In our case, the first ilr coordinate opposes the Left wing to the group of the Right wing and the Extreme Right party and the second opposes the Right wing to the Extreme Right party. The inverse ilr transformation is given by:

$$\mathbf{x} = \text{ilr}^{-1}(\mathbf{x}^*) = \mathcal{C}(\exp(V_D \mathbf{x}^*)) \text{ for } x^* \in \mathbb{R}^{D-1}$$

where the exponential of vector $\mathbf{x}$ is understood componentwise.

4.2 CODA regression models

In this paper, we use the notations in Table 2. Let $\mathbf{Y}_i$ denotes the compositional response value of the ith observation, $\mathbf{Y}_i \in \mathbf{S}^L$, and $\mathbf{X}_i^{(q)},\ q = 1, \ldots, Q$, denotes the value of the qth compositional covariate for the ith observation, $\mathbf{X}_i^{(q)} \in \mathbf{S}^{D_q},\ q = 1, \ldots, Q$, $Z_{ki},\ k = 1, \ldots, K$, denotes the kth classical covariate of the ith observation. Let us first introduce the CODA regression model in the ilr coordinate space as follows:

$$\text{ilr}(\mathbf{Y}_i) = \mathbf{b_0}^* + \sum_{q=1}^{Q} \text{ilr}(\mathbf{X}_i^{(q)}) \mathbf{B}_q^* + \sum_{k=1}^{K} Z_{ki} \mathbf{c}_k^* + \text{ilr}(\boldsymbol{\epsilon}_i) \tag{7}$$

Table 2: Notations

Variable	**Notation**	**Coordinates**
Dependent	$\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iL})$	$\text{ilr}(\mathbf{Y}_i) = \mathbf{Y}_i^*$
Compositional explanatory	$\mathbf{X}_i^{(q)} = (X_{i1}^{(q)}, \ldots, X_{iD_q}^{(q)})$	$\text{ilr}(\mathbf{X}_{ip}^{(q)}) = \mathbf{X}_{ip}^{(q)*}$
Classical explanatory	Z_{ki}	
General notations		
L	Number of components of the dependent variable	
$i = 1, \ldots, n$	Index of observations ($n = 95$)	
$q = 1, \ldots, Q$	Index of compositional explanatory variables ($Q = 3$)	
$p = 1, \ldots, D_q$	Index of the coordinates for the compositional explanatory variables	
$k = 1, \ldots, K$	Index of classical explanatory variables ($K = 5$)	

where $\text{ilr}(\mathbf{Y_i})$, $\text{ilr}(\mathbf{X}_i^{(q)})$ are the ilr coordinates of $\mathbf{Y}_i$, $\mathbf{X}_i^{(q)}$ $(q = 1, \ldots, Q)$ respectively, $\text{ilr}(\mathbf{Y_i}) \in \mathbb{R}^{L-1}$, $\text{ilr}(\mathbf{X}_i^{(q)}) \in \mathbb{R}^{D_q-1}$; $\mathbf{b_0}^*$, $\mathbf{B}_q^*$, $\mathbf{c}_k^*$ are the parameters in the coordinate space, and $\text{ilr}(\boldsymbol{\epsilon}_i)$ are the residuals in the coordinate space, $\text{ilr}(\boldsymbol{\epsilon}_i) \in \mathbb{R}^{L-1}$. The distributional assumption is $\text{ilr}(\boldsymbol{\epsilon})$ follows either the multivariate gaussian (N) distribution with with zero mean and covariance matrix $\boldsymbol{\Sigma}_N$ either the independent multivariate Student (IT) distribution with zero mean and covariance matrix $\boldsymbol{\Sigma}_{IT}$.
Let $\bigoplus$ denotes the summation, this regression model (7) can be written in the simplex as

$$\mathbf{Y}_i = \mathbf{b_0} \bigoplus_{q=1}^{Q} \mathbf{B}_q \boxdot \mathbf{X}_i^{(q)} \bigoplus_{k=1}^{K} Z_{ki} \odot \mathbf{c}_k \oplus \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n \tag{8}$$

where $\mathbf{b_0}, \mathbf{B}_1, \ldots, \mathbf{B}_Q, \mathbf{c}_1, \ldots, \mathbf{c}_K$ are the parameters satisfying $\mathbf{b_0} \in \mathbf{S}^L$, $\mathbf{B}_q \in \mathbf{S}^{D_q}$, $q = 1, \ldots, Q$, $\mathbf{c}_k \in \mathbf{S}^L$, $k = 1, \ldots, K$, $\mathbf{j}_L^T \mathbf{B}_q = \mathbf{0}_{D_q}$, $\mathbf{B}_q \mathbf{j}_{D_q} = \mathbf{0}_L$. The distributional assumption is that $\boldsymbol{\epsilon}_i \in \mathbf{S}^L$ follows either the multivariate gaussian (N) distribution (see Aitchison (1985)) either the independent multivariate Student (IT) distribution on the simplex .

We estimate the parameters of model (7) as in Section 3. Table 3 shows the estimated parameter for both case : Gaussian and Independent Student.

Table 3: Multivariate Gaussian and Student regression models with compositional and classical variables

	Gaussian model		*Student model, $\nu = 4$*	
	y_ilr[, 1]	y_ilr[, 2]	y_ilr[, 1]	y_ilr[, 2]
Constant	$+1.01(0.91)$	$-2.35(0.89)^{**}$	$+1.34(7.90)^{***}$	$-1.48(6.60)^{***}$
Age_ilr1	$+0.05(0.78)$	$-0.53(0.76)$	$+0.17(6.76)^{***}$	$+0.44(5.64)^{***}$
Age_ilr2	$-0.35(0.45)$	$-0.75(0.44)^{*}$	$-0.44(3.96)^{***}$	$-0.94(3.31)^{***}$
unemp_rate	$-7.31(2.77)^{**}$	$+13.1(2.71)^{***}$	$-7.94(24.1)^{***}$	$+10.6(20.1)^{*}$
income_tax_rate	$-0.42(1.00)$	$+0.19(0.98)$	$-1.02(8.69)^{***}$	$-0.82(7.26)^{***}$
Note:				*p<0.1; **p<0.05; ***p<0.01

 ISBN 978-84-947240-2-2

5 Model selections

Nguyen et al (2019) propose a a methodology to select a model between the Gaussian and independent Student models based on the Mahalanobis distance.

For an L-dimensional random vector $\mathbf{Y}$, with mean $\boldsymbol{\mu}$, and covariance matrix $\boldsymbol{\Sigma}$, the squared Mahalanobis distance is defined by:

$$d^2 = (\mathbf{Y} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{Y} - \boldsymbol{\mu})$$

If $\mathbf{Y}_1, \dots \mathbf{Y}_n$ is a sample of size n from the L-dimensional Gaussian distribution $\mathcal{N}_L(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)$, the squared Mahalanobis distance of observation i, denoted by d^2_{Ni}, follows a χ^2_L distribution. If $\boldsymbol{\mu}_N$ and $\boldsymbol{\Sigma}_N$ are unknown, then the squared Mahalanobis distance of observation i can be estimated by:

$$\hat{d}^2_{Ni} = (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_N)^T \hat{\boldsymbol{\Sigma}}_N^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_N)$$

where $\hat{\boldsymbol{\mu}}_N = \bar{\mathbf{Y}} = \dfrac{1}{n} \sum_{i=1}^{n} \mathbf{Y}_i$ and $\hat{\boldsymbol{\Sigma}}_N$ is the sample covariance matrix. This square distance follows a Beta distribution, up to a multiplicative constant:

$$\frac{n}{(n-1)^2} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_N)^T \hat{\boldsymbol{\Sigma}}_N^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_N) \sim Beta\left(\frac{L}{2}, \frac{n-L-1}{2}\right)$$

where L is the dimension of $\mathbf{Y}$. For large n, this Beta distribution can be approximated by the chi-square distribution $d^2_{Ni} \sim \chi^2_L$. According to Gnanadesikan (2011) (p. 172), $n = 25$ already provides a sufficiently large sample for this approximation, which is the case in all our examples below.

If we now assume that $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ is a sample of size n from the L-dimensional Student distribution $\mathbf{Y}_i \sim \mathbf{T}(\boldsymbol{\mu}_{IT}, \boldsymbol{\Sigma}_{IT}, \nu)$, then the squared Mahalanobis distance of observation i, denoted by d^2_{ITi} and properly scaled, follows a Fisher distribution (see Roth (2012)):

$$\frac{1}{L} \frac{\nu}{\nu - 2} d^2_{ITi} \sim \mathcal{F}(L, \nu)$$

If $\boldsymbol{\mu}_{IT}$ and $\boldsymbol{\Sigma}_{IT}$ are unknown, then the squared Mahalanobis distance of observation i can be estimated by:

$$\hat{d}^2_{ITi} = (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_{IT})^T \hat{\boldsymbol{\Sigma}}_{IT}^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_{IT}),$$

where $\hat{\boldsymbol{\mu}}_{IT}$ and $\hat{\boldsymbol{\Sigma}}_{IT}$ are the MLE of $\boldsymbol{\mu}_{IT}$ and $\boldsymbol{\Sigma}_{IT}$. Nguyen (2019) note that in the IT model, $\hat{\boldsymbol{\mu}}_{IT}$ is no longer equal to $\bar{\mathbf{Y}}$ and there is no result about the distribution of $\hat{d}^2_{ITi}$.

In the elliptical distribution family, the distribution of Mahalanobis distances characterizes the distribution of the observations. The merit of this approach is that the Mahalanobis distance is a one-dimensional variable. According to Nguyen et al (2019), we can test whether the Mahalanobis distances follow a chi-square distribution for testing the normality of the data. Similarly, we test whether the Mahalanobis distances follow the Fisher distribution for testing the Student distribution. As in Nguyen et al (2019), we perform some Kolmogorov−Smirnov tests in order to test different null hypothesis: Gaussian, Independent Student with three and four degrees of freedom.

Table 4 shows the p-values of these tests. At level 5%, we do reject the Gaussian distribution and we do not reject the Student distribution with three and four degrees of freedom.

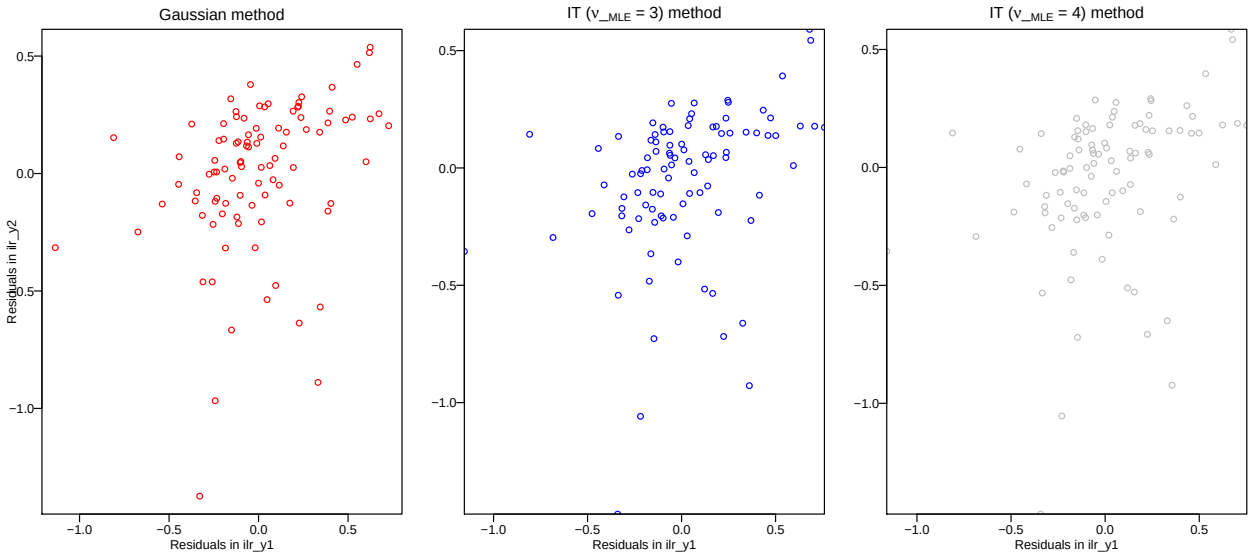


Figure 3: Scatterplots of residuals for the normal, IT ($\nu_{MLE} = 3$), and IT ($\nu_{MLE} = 4$) estimators

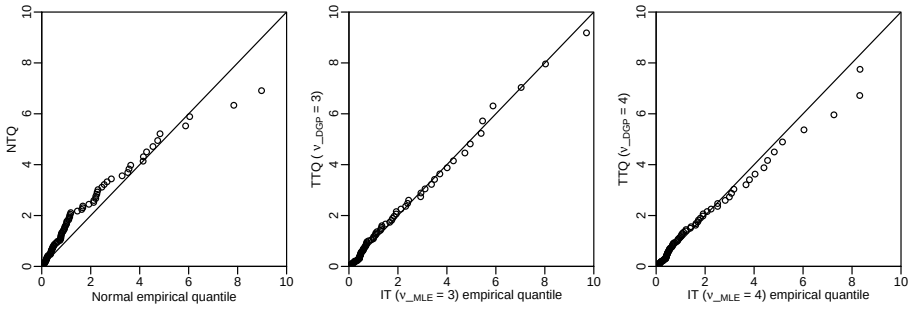


Figure 4: Q-Q plots of the Mahalanobis distances for the normal, IT ($\nu_{MLE} = 3$), and IT ($\nu_{MLE} = 4$) estimators

Figure 4 illustrate the Q-Q plots comparing the empirical quantiles of the Mahalanobis distances for the normal (respectively, the IT ($\nu_{MLE} = 3$), the IT ($\nu_{MLE} = 4$)) estimators on the horizontal axis to the theoretical quantiles of the Mahalanobis distances for the normal (respectively, the IT ($\nu_{MLE} = 3$), the IT ($\nu_{MLE} = 4$)) on the vertical axis. These Q-Q plots are coherent with the results of the tests in Table 4. The IT model with three degrees of freedom fits our data well.

6 Vote shares predictions

The interpretation of coefficients in regression model on the simplex is quite complex. Thus, in this section, we predict the vote share to understand how the socio-economics factors impact on the outcome of the election in France.

 ISBN 978-84-947240-2-2

Table 4: The p-values of the Mahalanobis distances tests with the null hypothesis and the corresponding estimators.

Hypothesis H_0 / **Method**	**P-values**
N	0.00
IT, $\nu_{MLE} = 3$	0.49
IT, $\nu_{MLE} = 4$	0.65

7 Conclusion

We have presented a CODA regression models on the simplex and compared two different models (multivariate Gaussian distribution and multivariate independent Student distribution). The results shows that the Student distribution is useful in the context of political economy. We have also predicted the vote shares to understand the impact of socio-economics factors on the departmental election in France.

Acknowledgment

This research project is within the scope of my PhD thesis work. I would like to thank Christine Thomas-Agnan and Anne Ruiz-Gazen for their supports and their helpful advices.

References

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

Aitchison, J. (1985). A general class of distributions on the simplex. *Journal of the Royal Statistical Society. Series B (Methodological)*, 136–146.

Ansolabehere, S. and W. Leblanc (2008). A spatial model of the relationship between seats and votes. *Mathematical and Computer Modelling 48*(9-10), 1409–1420.

Beauguitte, L. and C. Colange (2013). Analyser les comportements électoraux à l'échelle du bureau de vote. *Sciences de l'Homme et de la Société*.

Gnanadesikan, R. (2011). *Methods for statistical data analysis of multivariate observations*, Volume 321. John Wiley & Sons.

Honaker, J., J. N. Katz, and G. King (2002). A fast, easy, and efficient estimator for multiparty electoral data. *Political Analysis 10*(1), 84–100.

Johnson, N. L. and S. Kotz (1972). Distributions in statistics: Continuous multivariate distributions, john wiley & sons. *Inc. New York*.

Katz, J. N. and G. King (1999). A statistical model for multiparty electoral data. *American Political Science Review 93*(1), 15–32.

Kavanagh, A., S. Fotheringham, and M. Charlton (2006). A geographically weighted regression analysis of the election specific turnout behaviour in the republic of ireland. In *Elections, Public Opinion and Parties Conference, Nottingham 8th to 10th September*.

Kelejian, H. H. and I. R. Prucha (1985). Independent or uncorrelated disturbances in linear regression: An illustration of the difference. *Economics Letters 19*(1), 35–38.

Kotz, S. and S. Nadarajah (2004). *Multivariate t-distributions and their applications*. Cambridge University Press.

Mert, M. C., P. Filzmoser, G. Endel, and I. Wilbacher (2018). Compositional data analysis in epidemiology. *Statistical methods in medical research 27*(6), 1878–1891.

Morais, J., C. Thomas-Agnan, and M. Simionic (2017). Interpreting the impact of explanatory variables in compositional models. *TSE Working Paper*.

Nguyen, T. H. A., T. Laurent, C. Thomas-Agnan, and A. Ruiz-Gazen (2018). Analyzing the impacts of socio-economic factors on french departmental elections with coda methods. *TSE Working paper*.

Nguyen, T. H. A., A. Ruiz-Gazen, T. Laurent, and C. Thomas-Agnan (2019). Multivariate Student versus multivariate Gaussian regression models with application to finance. *Special Issue "Applied Econometrics", Journal of Risk Financial Management*.

Prucha, I. R. and H. H. Kelejian (1984). The structure of simultaneous equation estimators: A generalization towards nonnormal disturbances. *Econometrica: Journal of the Econometric Society*, 721–736.

Seber, G. A. (2009). *Multivariate observations*, Volume 252. John Wiley & Sons.

Independence test for compositional tables

V. Pawlowsky-Glahn[1], M. Planes-Pedra[1], and J.J. Egozcue[2]
[1]University of Girona, Girona, Spain; *vera.pawlowsky@udg.edu*; *montserrat.planes@udg.edu*
[2]Technical University of Catalonia, Barcelona, Spain; *juan.jose.egozcue@upc.edu*

Summary

Contingency tables can be decomposed in independence and interaction tables. Given a sample of compositional tables, the need for a test arises to answer the question whether the mean table is an independent table, or, on the contrary, there are significant interactions. With this purpose we develop a bootstrap test to check the hypothesis that the interaction table is the neutral element, that is, that all entries of the interaction table are equal. It is based on the idea (a) to generate via bootstrap a sample of independent tables, (b) to compute for each independent table its Aitchison norm, (c) to build the empirical distribution of the norms, and (d) to compare the norm of the original table with the obtained distribution. Survey data, corresponding to 145 undergraduate students (76% females, 24% males) asked to evaluate three protective measures in 10 different situations, and annual mortality rates in some European countries are used for illustration.

Key words: compositional tables, interaction and independence tables, bootstrap test.

1 Introduction

Compositional tables (Fačevicová and Hron, 2015; Fačevicová et al., 2016) can be uniquely decomposed into orthogonal independent and interaction tables (Egozcue et al., 2015). Let X be a composition arranged in a (D_1, D_2)-matrix, then

$$X = X_{\text{ind}} \oplus X_{\text{int}}\,, \qquad \langle X_{\text{ind}}, X_{\text{int}} \rangle_a = 0\,, \tag{1}$$

where both X_{ind} and X_{int} are (D_1, D_2)-matrices; $\oplus$ is the perturbation of compositions shaped as matrices; and $\langle \cdot, \cdot \rangle_a$ stands for the Aitchison inner product of compositions also shaped as a matrix (Pawlowsky-Glahn and Egozcue, 2001; Egozcue et al., 2015; Ortego and Egozcue, 2016). The independent part X_{ind} is obtained as the perturbation of the geometric marginals of X. They are tables of equal rows, respectively columns, obtained as the geometric mean by rows, respectively by columns. The table X_{ind} is then the closest independent table to X in the Aitchison distance sense. The interaction table is obtained by perturbation subtraction $X_{\text{int}} = X \ominus X_{\text{ind}}$. Note that the removal of the independent table from X is equivalent to a double compositional centring of X.

When an n-sample of compositional tables is available, the study of the sample compositional mean, also known as sample center,

$$\bar{X} = \frac{1}{n} \odot \bigoplus_{i=1}^{n} X_i\,,$$

can be of interest (Pawlowsky-Glahn et al., 2017, 2019). The orthogonal decomposition (1) allows to analyse the mean compositional table of a sample of compositional tables and poses the question whether the empirical mean table can be considered independent or, on the contrary, this hypothesis has to be rejected.

ISBN 978-84-947240-2-2

Consequently, a test on the mentioned hypothesis is needed. Up to our knowledge, no such test is available. In the present contribution we propose a bootstrap test, which is described in Section 2. The test is illustrated with data from a survey sample in Section 3 and, in Section 4, using mortality rates in some countries of the European Union.

2 Independence test

Given a sample of (D_1, D_2) compositional tables, $X_i, i = 1, 2, ...n$, the null hypothesis of interest to be tested is that the mean table, $\bar{X}$, is independent or, equivalently, that the mean interaction table is the neutral element, $\bar{X}_{\text{int}} = N$, against the alternative hypothesis that the mean interaction table is not negligible, i.e., $\bar{X}_{\text{int}} \neq N$. Formally,

$$\mathcal{H}_0 : \bar{X} = \bar{X}_{\text{ind}}\,, \quad \mathcal{H}_1 : \bar{X} \neq \bar{X}_{\text{ind}}\,,$$

or, equivalently,

$$\mathcal{H}_0 : \bar{X}_{\text{int}} = N\,, \quad \mathcal{H}_1 : \bar{X}_{\text{int}} \neq N\,.$$

To perform this test, m bootstrap samples are generated from the original sample $X_i, i = 1, 2, ...n$. For each re-sample the compositional mean table, $\bar{X}_\ell$, $\ell = 1, 2, ..., m$, is computed. The m tables $\bar{X}_\ell$ are decomposed into their independent and interaction parts. Then, for each of the independent tables, $\bar{X}_{\ell,\text{ind}} = [x_{ij\ell}]$, the square Aitchison norm

$$\|\bar{X}_{\ell,\text{ind}}\|_a^2 = \frac{1}{2D}\sum_{i=1}^{D_1}\sum_{j=1}^{D_2}\sum_{i'=1}^{D_1}\sum_{j'=1}^{D_2}\left(\ln\frac{x_{ij\ell}}{x_{i'j'\ell}}\right)^2, \qquad D = D_1 \times D_2\,,$$

is computed. The square norm of the sample mean table $\|\bar{X}\|_a^2$ is then compared to the empirical distribution of the norms $\|\bar{X}_{\ell,\text{ind}}\|_a^2$, $\ell = 1, 2, \ldots, m$. This empirical cumulative distribution is denoted F^*, where the asterisk recalls its empirical character. The corresponding p-value is $\alpha_p = 1 - F^*(\|\bar{X}\|_a^2)$, and, for a given significance level α such that $\alpha_p \leq \alpha$, the hypothesis $\mathcal{H}_0$ must be rejected.

3 A survey on contraceptive measures

In a survey $n = 145$ undergraduate students (76% females, 24% males) were asked to evaluate three protective measures (P preservative, C contraceptive pill, M morning after pill) during sexual intercourse. Ten different situations (called items) were considered (Pawlowsky-Glahn et al., 2017, 2019), and the students were asked to evaluate the three protective measures in each of the situations. The scores for (P,C,M) in each item had to be positive and such that they added up to 100. Therefore, the minimum score, for an item and for a measure, was 1 and the maximum 98. Incorrect scores were modified to fit the standards; for instance, a score equal to 0 was changed to 1. This survey generated a sample, with $n = 145$, of $D_1 \times D_2 = 3 \times 10$ compositional tables (described in Table 1) in a closed form. The mean table and its decomposition, is

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \\ 10 \end{array}\!\!\begin{array}{c} \begin{array}{ccc} P & C & M \end{array} \\ \begin{pmatrix} 0.1047 & 0.0054 & 0.0032 \\ 0.0513 & 0.0420 & 0.0109 \\ 0.0487 & 0.0454 & 0.0066 \\ 0.0561 & 0.0277 & 0.0134 \\ 0.1105 & 0.0040 & 0.0024 \\ 0.0882 & 0.0107 & 0.0038 \\ 0.0077 & 0.0674 & 0.0192 \\ 0.0100 & 0.0663 & 0.0168 \\ 0.0405 & 0.0347 & 0.0151 \\ 0.0744 & 0.0077 & 0.0053 \end{pmatrix} \end{array} = \begin{pmatrix} 0.0366 & 0.0169 & 0.0063 \\ 0.0860 & 0.0396 & 0.0147 \\ 0.0732 & 0.0338 & 0.0126 \\ 0.0825 & 0.0380 & 0.0141 \\ 0.0304 & 0.0140 & 0.0052 \\ 0.0460 & 0.0212 & 0.0079 \\ 0.0645 & 0.0297 & 0.0111 \\ 0.0671 & 0.0309 & 0.0115 \\ 0.0830 & 0.0383 & 0.0142 \\ 0.0433 & 0.0199 & 0.0074 \end{pmatrix} \oplus \begin{pmatrix} 0.0910 & 0.0101 & 0.0164 \\ 0.0190 & 0.0338 & 0.0236 \\ 0.0212 & 0.0428 & 0.0167 \\ 0.0216 & 0.0232 & 0.0301 \\ 0.1156 & 0.0090 & 0.0146 \\ 0.0611 & 0.0161 & 0.0154 \\ 0.0038 & 0.0722 & 0.0553 \\ 0.0048 & 0.0683 & 0.0466 \\ 0.0155 & 0.0289 & 0.0338 \\ 0.0548 & 0.0123 & 0.0226 \end{pmatrix}, \tag{2}$$

 ISBN 978-84-947240-2-2

Table 1: Survey data as a compositional table. P = preservative; C = contraceptive pills; M = morning after pill.

item	**protective measure**	P	C	M
1	protect from sexually transmitted infections (STI)	P1	C1	M1
2	protect from pregnancy	P2	C2	M2
3	provide peace during and after intercourse	P3	C3	M3
4	economically accessible	P4	C4	M4
5	protect from the transmission of the AIDS virus	P5	C5	M5
6	evidence interest in protecting the health of the couple	P6	C6	M6
7	increase feelings of pleasure in man	P7	C7	M7
8	increase feelings of pleasure in woman	P8	C8	M8
9	are easy to use correctly	P9	C9	M9
10	do not cause side effects	P10	C10	M10

where the first table in the right hand member is $\bar{X}_{\text{ind}}$ and the second is $\bar{X}_{\text{int}}$. The main goal of interest in a survey like the one presented, is the interaction table, $\bar{X}_{\text{int}}$, provided that it is non-neutral. If $\bar{X}_{\text{int}} = N$, it would mean that the preventive measures are evaluated independently of the items, and no relevant information is then obtained.

The interpretation of interactions in $\bar{X}_{\text{int}}$ is easier when based on $\text{clr}(\text{Cen}(X_{\text{int}}))$ (see Figure 1). The squares

	P	C	M
1	1.3	−0.89	−0.41
2	−0.26	0.31	−0.05
3	−0.16	0.55	−0.39
4	−0.13	−0.06	0.2
5	1.54	−1.01	−0.53
6	0.9	−0.43	−0.48
7	−1.87	1.07	0.8
8	−1.65	1.02	0.63
9	−0.47	0.15	0.31
10	0.79	−0.7	−0.09

Figure 1: Overall interaction table (clr). Colors enhance importance of interaction (I). Red: strong positive I. Ocre: medium positive I. Grey: low positive I. White: no I. Pale blue: low negative I. Aquamarine: medium negative I. Dark blue: strong negative I.

of the entries add up to the mean simplicial deviance (Egozcue et al., 2015), which is a measure of cross information between methods and items. The simplicial deviance is the square Aitchison distance to the neutral element N which represents the independence. Positive (negative) entries point at cells in which the score is larger (less) than that predicted by the independent table. Note that the sum of entries of the matrix $\text{clr}(\bar{X}_{\text{int}})$ is zero by rows and columns; therefore, any positive entry should be compensated by negative entries in the same row and column. For a more detailed interpretation of Table 1 see Pawlowsky-Glahn et al. (2019).

To answer the question whether the mean table is an independent table, or if, on the contrary, there are significant interactions, the above described bootstrap test was conducted. Figure 2 shows the bootstrap

distribution of $\|\bar{X}_{\text{ind}}\|_a^2$ compared with the observed $\|\bar{X}\|_a^2$. The obtained p-value is less than 10^{-4}. The simplicial deviance of $\bar{X}$ is equal to $\|\bar{X}_{\text{int}}\|_a^2 = 19.29$, and its relative value is $\|\bar{X}_{\text{int}}\|_a^2/\|\bar{X}\|_a^2 = 0.498$, this is, the square Aitchison norm of the interaction $\bar{X}_{\text{int}}$ is similar to that of $\bar{X}_{\text{ind}}$, thus confirming the importance of the interaction and the lack of independence.

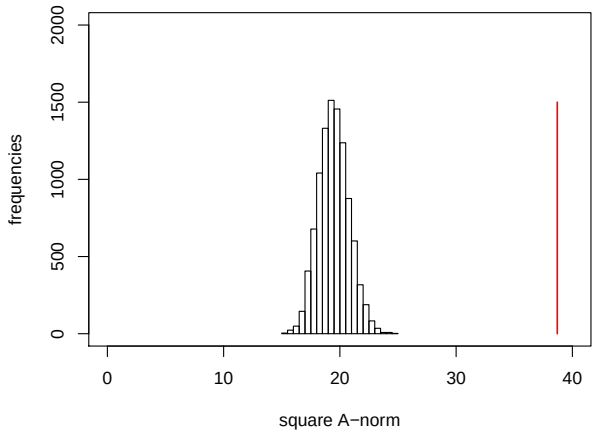


Figure 2: Histogram of the square Aitchison norm of independent mean tables obtained in a bootstrap m-sample ($m = 10000$), centered on 19.29. The red line is placed at the square Aitchison norm of the observed $\bar{X}$, which is 38.70, quite apart from independence (p-value less than 10^{-4}).

4 Mortality rates

Annual mortality rates of European countries estimated by causes of death are available in Eurostats (2019). From this data base a small group of data were selected: only females; age under 65; some large countries in Europe (Germany -DE-, Spain -ES-, France -FR-, Italy -IT-, United Kingdom -UK-); only some causes of death (cancer -Can-, circulatory causes -Cir-, digestive causes -Dig-, HIV related, respiratory causes -Res-, skin deseases -Ski-). The tables country/cause of death were available for 5 years 2011-2015; they are considered as table samples ($n = 5$).

The question to be examined is whether the causes of death depend on the country or not. The test explained in Section 2 was conducted in this case with only $m = 500$ re-samples. This reduced number of bootstrap re-samples is adapted to the fact that the original number of samples (years) is only 5. The results are shown in Figure 3. In the left panel, the histogram of $\|\bar{X}_{\ell,\text{ind}}\|_a^2$ is compared with the sample value of $\|\bar{X}\|_a^2 = 138.4$ which is out of the range of the histogram (p-value less than 0.01). The relative simplicial deviance is 0.031, a very small value indicating that the interaction square norm is a small fraction of the total square norm, although independence is significantly rejected. The right panel of Figure 3 shows the clr-interaction matrix. In the clr-interaction table the HIV column shows the highest values of interactions; also the row corresponding to UK is showing considerable interactions. Paying attention to HIV we realize that the mortality rates in ES and IT are large in contrast with DE, FR, and specially with UK, where there is a relevant relative deficit of mortality rate in HIV. In turn, other causes of mortality in UK are higher than the predicted by the independent table, specially those of Res and Ski. These last interactions may be due to climate characteristics of the countries.

In order to show a case where independence of the mean table is not rejected in the test, a subtable o the

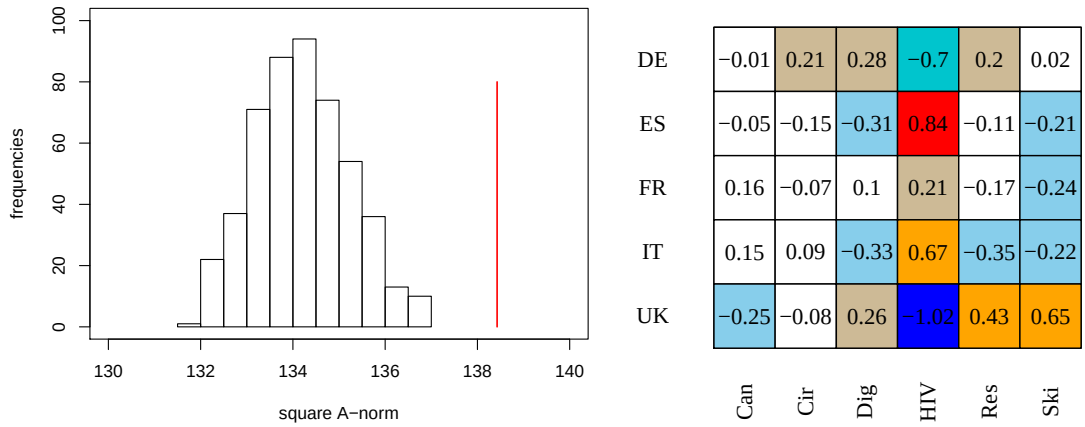


Figure 3: Left panel: histogram of 500 bootstrap re-samples of $\|\bar{X}_{\ell,\text{ind}}\|_a^2$ compared with the sample value of $\|\bar{X}\|_a^2$. The plot suggests a cautious rejection of the independence of $\bar{X}$ reflected in the small relative simplicial deviance 0.031. Right panel: clr-interaction matrix $\bar{X}_{\text{int}}$. Stronger colors point out the cells with larger interaction.

previous one has been selected. The columns HIV and Skin, and the row corresponding to UK, where strong interactions are concentrated (Figure 3) have been suppressed. Note that clr-interaction values change when taking a subcomposition (a subtable in this case). In this reduced case, the analysis results in a large p-value equal to 0.145, which does not suggest a rejection of independence (Fig. 4, left panel). Also $\|\bar{X}\|_a^2 = 18.88$ (red line) is placed within the histogram of $\|\bar{X}_{\ell,\text{ind}}\|_a^2$ as a confirmation of the obtained p-value. The right panel of Figure 4 shows a moderate uniform spread of interactions, not far from the neutral element N, reflected in a low relative simplicial deviance (0.015).

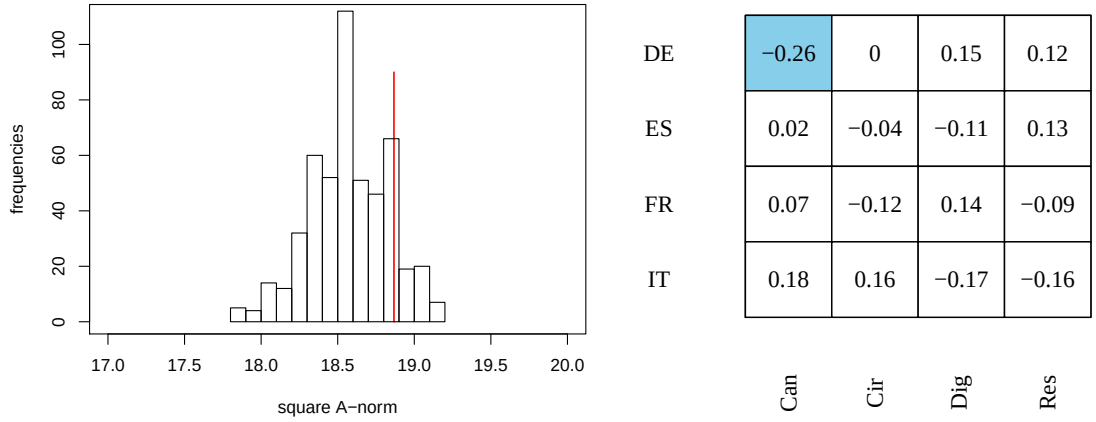


Figure 4: Left panel: histogram of 500 bootstrap re-samples of $\|\bar{X}_{\ell,\text{ind}}\|_a^2$ compared with the sample value of $\|\bar{X}\|_a^2$. The plot suggests a cautious rejection of the independence of $\bar{X}$ reflected in the small relative simplicial deviance 0.031. Right panel: clr-interaction matrix $\bar{X}_{\text{int}}$. Stronger colors point out the cells with larger interaction.

5 Conclusions

A bootstrap test of hypothesis on the independence of the mean compositional table from a sample has been designed. It is based on the orthogonal decomposition of a compositional table into its independent and interaction parts. The presented examples demonstrate its performance.

Acknowledgements

This research has received financial support from the Spanish Ministry of Education and Science under project 'CODA-RETOS' (Ref. MTM2015-65016-C2-1 (2)-R (MINECO/FEDER,UE)).

References

Egozcue, J. J., V. Pawlowsky-Glahn, M. Templ, and K. Hron (2015). Independence in contingency tables using simplicial geometry. *Communications in Statistics –Theory and Methods 44*(18), 3978–3996.

Eurostats (2019). Causes of death - standardised death rate by residence. Downloaded March 16, 2019. https://data.europa.eu/euodp/es/data/dataset/vL1BRcXY3uId3Ecuu4d21g.

Fačevicová, K. and K. Hron (2015). Covariance structure of compositional tables. *Austrian Journal of Statistics 44*(3), 31–44.

Fačevicová, K., K. Hron, V. Todorov, and M. Templ (2016). General approach to coordinate representation of compositional tables. *Scandinavian Journal of Statistics 43*(4), 962–977.

Ortego, M. I. and J. J. Egozcue (2016). Bayesian estimation of the orthogonal decomposition of a contingency table. *Austrian Journal of Statistics 45*(4), 45–56.

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

Pawlowsky-Glahn, V., J. J. Egozcue, and M. Planes-Pedra (2017). Survey data on perceptions of contraceptive measures as compositional tables. In K. Hron and R. Tolosana-Delgado (Eds.), *Proceedings of CoDaWork2017*, http://www.coda-association.org/en/, pp. 229–304.

Pawlowsky-Glahn, V., J. J. Egozcue, and M. Planes-Pedra (2019). Survey data on perceptions of contraceptive measures as compositional tables. *Revista Latinoamericana de Psicología*. In press.

Joint evolution of access to water of urban and rural populations in South America through Compositional Data Analysis.

F.A. Quispe-Coica[1] and A. Pérez-Foguet[1]
[1]Research Group on Engineering Sciences and Global Development, Dept. of Civil and Environmental Engineering, School of Civil Engineering, Universitat Politècnica de Catalunya BarcelonaTech, Barcelona, Spain.
filimon.alejandro.quispe@upc.edu; agusti.perez@upc.edu

Summary

In the international water and sanitation sector, it is usual to use multivariate statistical methods to monitor and report the overall progress of access to water and sanitation. However, the methods used do not take into account the compositional characteristics of the data. Recently, to overcome this problem, the application of multivariate temporal interpolation models that include this characteristic has been proposed.

On the other hand, it is usual to carry out analyzes separately between the urban and rural sectors (WHO/UNICEF, 2015), even though both parties form a whole (the general population of the country). The disaggregated work does not allow to see the crossed influence between the population evolution and the levels of service. In addition, according to (Cohen, 2004), the question arises whether disaggregated work is a good option or not, given that the definition of both categories is not homogeneous, which makes international comparisons difficult.

Therefore, this work focuses on the comparative analysis between the joint and disaggregated treatment of indicators of access to water, in urban and rural contexts, considering that the level of service is divided into four categories. Different temporal interpolation models are applied to the balances defined between the parties through the isometric logratio (ilr) transformation (Egozcue et al. 2003). The identification of outliers is included through Mahalanobis distance.

The preliminary results obtained with the data from the countries of South America show that the adjusted models may depend on the joint or disaggregated approach. The quality metrics Nash-Sutcliffe Efficiency (NSE) and Root Mean Square Error (RMSE) confirm that the best adjustments are obtained in one way or another depending on the case. The outliers are different in each situation.

Key words: CoDa aggregate and disaggregate, models, water, outliers.

ISBN 978-84-947240-2-2

1 Introduction

In the period (Year: 2000-2015) of the Millennium Development Goals (MDGs), the global monitoring and reporting of access to water and sanitation has been in charge of the Joint Monitoring Program (JMP), between the World Health Organization (WHO) and United Nations International Children's Emergency Fund (UNICEF). After 2015, in the context of the Sustainable Development Goals (SDG) (Year: 2015-2030), the JMP continues to make the global reports (WHO/UNICEF, 2017).

For the preparation of the report, the main sources of information are the household survey, the census, and the administrative data compiled by governmental and non-governmental entities (JMP, 2018). With this information, estimates are made separately for the urban and rural sector, while national estimates are generated as the weighted average of the two using population data (JMP, 2018). Due to this, the joint evolution in the time series is not visualized, nor the cross-influence between the evolution of the population and the levels of service.

On the other hand, the definitions of the urban and rural are in question, when they differ in many countries of Latin America and the Caribbean (Dirven et al., 2011). Therefore, the JMP is forced to rely on the existing definitions of the member countries of the SDGs. A clear example of this is Peru, in which rural population is considered to be one that does not exceed 2,000 thousand inhabitants (DS. N°031-2008-VIVIENDA, 2008); Meanwhile, in Chile it is considered rural, *a human settlement with a population less than or equal to 1,000 inhabitants, or between 1,001 and 2,000 inhabitants where more than 50% of the population that declares to have worked is dedicated to primary activities* (INE, 2018). It is in this scenario that the question arises whether disaggregated work is a good option or not, given that the definition of both categories is not homogeneous, which makes international comparisons difficult(Cohen, 2004).

Another situation is presented in the data used for national, regional and global estimates; when it is appreciated that water and sanitation service levels have been disaggregated into four parts for the urban and rural sector. Being these, the services improved (Piped on premises, Other improved) and unimproved (Surface water, Other unimproved), for the case of water. While for sanitation they are: Improved, Shared, Open defecation, Unimproved. Also called water and sanitation ladders (WHO/UNICEF, 2015). What shows that these data have compositional characteristics, because they are part of a whole, and the sum is a constant (Aitchison, 1986). Post-2015, they remain compositional, but in five water and sanitation ladders (WHO/UNICEF, 2017). Therefore, they have to be addressed as such. For this, there are already antecedents in the sector that show the importance of taking into account the compositional nature of the population data, when the estimates of access to water sanitation and hygiene (WASH) are modeled (Pérez-Foguet et al., 2017). In addition, the statistical analysis in the world of compositional data (CoDa) is a good alternative to evaluate and visualize the aggregate information of the urban and rural sector, in the time series.

This new method in the sector follows statistical procedures of compositional data, for which, certain disadvantages must be overcome, one of them being the presence of zero values; because the transformations carry proportions in which this is not possible. To address this, in literature there are different methods of work (Josep Antoni Martín-Fernández et al., 2011; Palarea-Albaladejo et al., 2008; J. A. Martín-Fernández et al., 2003; Templ et al., 2016).

Therefore, the purpose of this study is to assess the cross-influence between the evolution of the population and service levels; value whether disaggregated work is a good option or not. For which, statistical methods for compositional data will be used. In addition, the presence of outliers in the models will be analyzed, with the purpose of assessing in which of them are generated mostly. The hypothesis that estimates can vary in certain situations if they have different SBPs will be tested. The generalized additive model (GAM) will be used to generate and compare the data.

The article will be organized as follows: The following section presents the work methodology in aggregate and disaggregated data of the sector. In section 3, the results are analyzed and discussed. In section 4, the conclusions of this work and the future challenges in the WASH sector are described.

ISBN 978-84-947240-2-2

2 Materials and Method

The information used for the study is available on the JMP platform (JMP, 2017). In this there are data on access to water, sanitation and hygiene (WASH), ordered by sector (Urban and Rural).

2.1 Selection of countries and indicators.

The countries in the analysis are: Bolivia, Colombia, Ecuador, Paraguay, Peru and Uruguay. They respond to water access indicators with compositional characteristics, with a minimum data quantity of twelve (Uruguay) and a maximum of twenty-six (Peru). Other countries are not considered for lack of one or more variables that prevent the formation of compositions. In addition, they do not comply with the minimum amount of data to use GAM (Fuller et al., 2016). As for the population, they are the same ones with which the JMP makes the estimates of each country. In this is the population divided into urban and rural.

The indicators of analysis in the period of the MDGs have been related to the monitoring of the population that accesses improved and unimproved water, made up of water and sanitation ladders. In this study, we followed the same classification and used the same population data from JMP for estimates of access to services, according to the indicator.

The services of access to improved and unimproved water are classified as follows:

Access to improved water: They are supplied by networks and other improved forms.

- **Piped water** (Xu,r_1)**:** They are considered like this, the access of the water by public network inside the house, public network outside the house but inside the building, public tap and others.
- **Other improved sources** (Xu,r_2)**:** Tank truck, and other forms of access to improved water that is not piped.

Access to unimproved water: They are supplied from surface sources and other unimproved sources.

- **Surface water** (Xu,r_3)**:** According to the country they can be, river, spring, irrigation channel, and other.
- **Other unimproved sources** (Xu,r_4)**:** Other non-surface water sources

The disaggregated analysis has a composition of four parts for the urban sector and four for the rural sector. While the aggregate analysis between urban and rural, they carry compositions of eight parts. These are represented as follows:

$$x_{r\_1} + x_{r\_2} + x_{r\_3} + x_{r\_4} = 1 \quad \text{Eq. (1)}$$

$$x_{u\_1} + x_{u\_2} + x_{u\_3} + x_{u\_4} = 1 \quad \text{Eq. (2)}$$

$$x_{r\_1} + x_{r\_2} + x_{r\_3} + x_{r\_4} + x_{u\_5} + x_{u\_6} + x_{u\_7} + x_{u\_8} = 1 \quad \text{Eq. (3)}$$

Where: Xu,r_1= Piped water; Xu,r_2=Other improved sources; Xu,r_3 =Surface water; Xu,r_4 = Other unimproved sources.
*Xu,r: Urban or rural value.

To make CoDa in eight parts, the population of the urban and rural sector has been multiplied with the proportions in CoDa of four parts (Eq. (1) y Eq. (2)), as appropriate. Then, the population expressed in eight parts was divided among the total population. Thus forming the compositions shown in Eq. (3).

2.2 Balances and transforms

Balances are defined in a group of parties that have access to improved and unimproved services.

ISBN 978-84-947240-2-2

Given this premise, there is a need to divide the aggregate analysis (eight-part CoDa) into two scenarios (Figure 1-*V3* y Figure 2-*V4*). The first will be when the main proportion is defined by sectors (proportion of people who have access to water in the rural sector among people who access water in the urban sector (see *V3*)). After this, the proportion criterion between improved and unimproved services is applied internally. In the second scenario, balances are made between access to improved and unimproved services, in the total (see *V4*); unlike the previous one, the group of parts is not separated by sector. Regarding CoDa of four parts, in both scenarios the same balance is maintained (*V1* and *V2*)

Next, balance "V" is made according to Egozcue and Pawlowsky-Glahn, (2005).

Scenario 1: Balance of eight parts, is carried out by sectors (Rural and Urban).

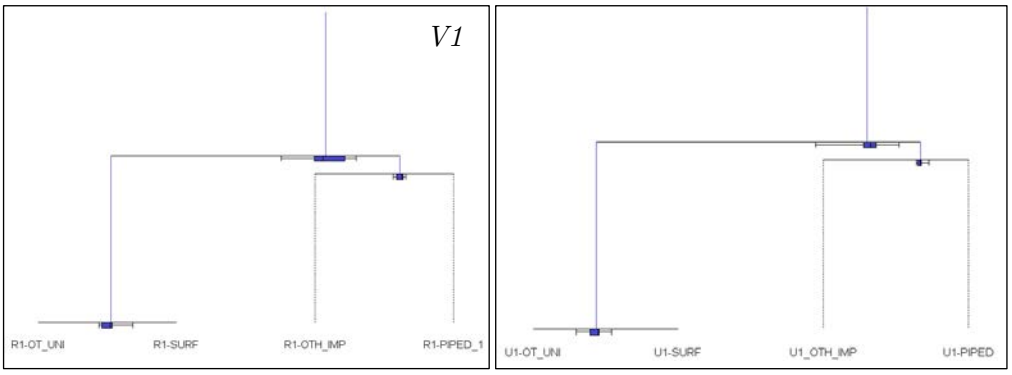
- Rural: Group of parts between the improved (Xr_1, Xr_2) and the unimproved (Xr_3, Xr_4). Balance *V1*, shows the partitions.
- Urban: Group of parts between the improved (Xu_1, Xu_2) and the unimproved (Xu_3, Xu_4). Balance *V2*, shows the partitions.
- Aggregate data urban and rural: Group of parts between rural (Xr_1, Xr_2, Xr_3, Xr_4) and urban (Xu_5, Xu_6, Xu_7, Xu_8). Internally, they will be classified as a group of access to improved and unimproved water. Balance *V3*, shows the partitions.

$$V_1 = \begin{pmatrix} x_{r1} & x_{r2} & x_{r3} & x_{r4} \\ +1 & +1 & -1 & -1 \\ +1 & -1 & 0 & 0 \\ 0 & 0 & +1 & -1 \end{pmatrix}$$

$$V_2 = \begin{pmatrix} x_{u1} & x_{u2} & x_{u3} & x_{u4} \\ +1 & +1 & -1 & -1 \\ +1 & -1 & 0 & 0 \\ 0 & 0 & +1 & -1 \end{pmatrix}$$

$$V_3 = \begin{pmatrix} x_{r1} & x_{r2} & x_{r3} & x_{r4} & x_{u5} & x_{u6} & x_{u7} & x_{u8} \\ +1 & +1 & +1 & +1 & -1 & -1 & -1 & -1 \\ +1 & +1 & -1 & -1 & 0 & 0 & 0 & 0 \\ +1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & +1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & +1 & +1 & -1 & -1 \\ 0 & 0 & 0 & 0 & +1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & +1 & -1 \end{pmatrix}$$

V2



V3

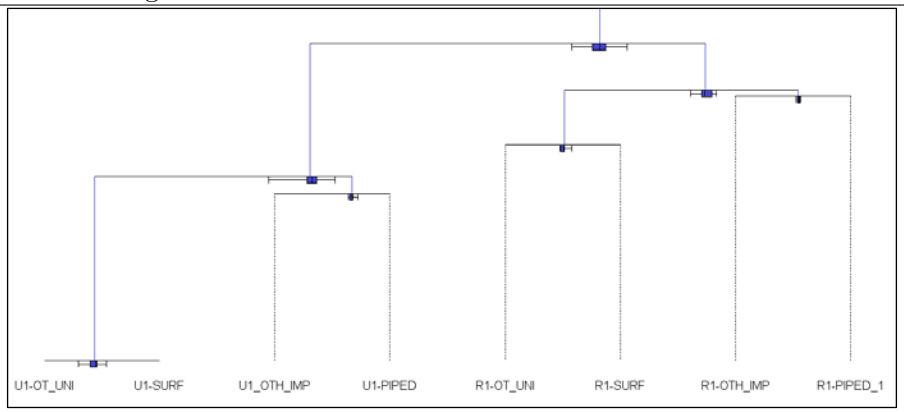


Figure 1: Balance of Colombia – Scenario 1.

Scenario 2: Balance of eight parts is made between access to improved water and not improved, in the total of variables.

- Rural: Group of parts between the improved (Xr_1, Xr_2) and the unimproved (Xr_3, Xr_4). Balance *V1*, shows the partitions.
- Urban: Group of parts between the improved (Xu_1, Xu_2) and the unimproved (Xu_3, Xu_4). Balance *V2*, shows the partitions.
- Aggregate data urban and rural: Group of parts between rural (Xr_1, Xr_2, Xu_5, Xu_6) and urban (Xr_3, Xr_4, Xr_7, Xr_8). Internally, they will be classified as a group of access to improved and unimproved water. Balance *V4*, shows the partitions.

$$V_1 = \begin{pmatrix} x_{r1} & x_{r2} & x_{r3} & x_{r4} \\ +1 & +1 & -1 & -1 \\ +1 & -1 & 0 & 0 \\ 0 & 0 & +1 & -1 \end{pmatrix} \qquad V_2 = \begin{pmatrix} x_{u1} & x_{u2} & x_{u3} & x_{u4} \\ +1 & +1 & -1 & -1 \\ +1 & -1 & 0 & 0 \\ 0 & 0 & +1 & -1 \end{pmatrix}$$

$$V_4 = \begin{pmatrix} x_{r1} & x_{r2} & x_{r3} & x_{r4} & x_{u5} & x_{u6} & x_{u7} & x_{u8} \\ +1 & +1 & -1 & -1 & +1 & +1 & -1 & -1 \\ +1 & +1 & 0 & 0 & -1 & -1 & 0 & 0 \\ +1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & +1 & -1 & 0 & 0 \\ 0 & 0 & +1 & +1 & 0 & 0 & -1 & -1 \\ 0 & 0 & +1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & +1 & -1 \end{pmatrix}$$

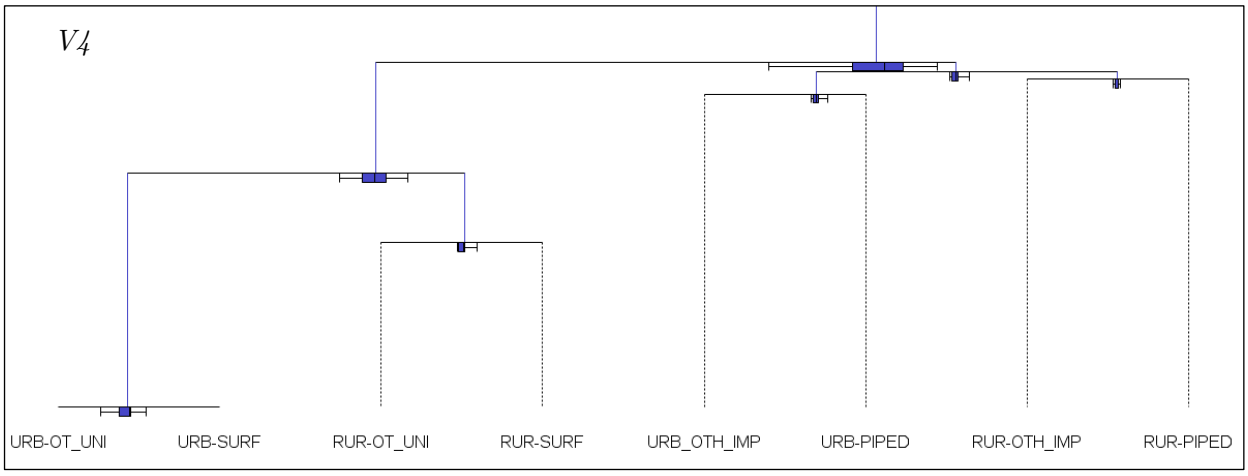


Figure 2: Balance of Colombia – Scenario 2.

After balances, isometric log-ratio transformations are performed (J. J. Egozcue et al., 2003). For this, the following equation is used.

ISBN 978-84-947240-2-2

$$Y^t = ilr = \sqrt{\frac{r \times s}{r+s}} \ln \frac{g_m(X_r+)}{g_m(X_s-)} \quad \text{Eq. (4)}$$

r = Number of positive variables in the balance V.

s = Number of negative variables in the balance V.

$g_m(\text{-})$ = It is the geometric mean of the variables.

2.3 Metrics of analysis

The existence of outliers is detected using the robust Mahalanobis distance (Filzmoser et al., 2008). Subsequently, it is compared according to the amount of data that are considered outliers in four-part CoDa and eight-part CoDa. Preliminary analysis removing outliers, show variation in estimates with aggregate and disaggregated data. Therefore, the comparison between eight-part CoDa and four-part CoDa is made without removing these values. This in order that you are comparatives are carried out under similar conditions.

For time series estimates, GAM (k = 4) is used. This is supported by the presence of non-linear data (Fuller et al., 2016). In addition, the flexibility of GAM helps to better adapt to the data in the time series.

Estimation results are compared between four-part CoDa (urban and rural) and eight-part CoDa. For this, the values estimated in eight-part CoDa are multiplied by the total population (the same population used to make eight-part CoDa in item 2.1) and then disaggregated into four-part CoDa for urban and rural. The result of these is compared with the estimates made in a disaggregated manner. This comparison is made with RMSE metric. If the value is zero, then the values estimated in eight-part CoDa generate same values when working independently in rural and urban areas (four-part CoDa). Otherwise, estimated values differ between the two. For the latter, the predictive capacity between CoDa 8 and CoDa 4 is evaluated and compared. For this, the NSE indicators are used and comparisons are made with the observed values in the rural and urban sectors (Table 3).

All this has been carried out and implemented in R Core Team (2018), using the following statistical packages: **nlme, compositions and mgcv**, by Pinheiro et al., (2018); Boogaart et al., (2014) & Wood, (2017), respectively. To treat data with outliers in CoDa, the **robCompositions** statistical package was used (Templ et al., 2011).

3 Results and discussions

3.1 ILR transformations and outlier

Scenario 1 and scenario 2, presented the same number of outliers. According to SBP raised in Scenario 1, it shows that metrics of R-adj and the ilr of Figure 3A, and 3B are similar to six transforms of Figure 3C (ilr2, ilr3, ilr4, ilr5, ilr6 y ilr7). This occurs because they maintain proportions of CoDa 4. The only one that changes is the ilr1 of Figure 3C, because it is represented with proportions between rural and urban water (*V3*). This can be a factor so that estimates in aggregated and disaggregated data do not have significant variation, as shown in Figure 4G, 4H. On the other hand, this does not occur when the balance (*V4*) varies (Figure 4J, 4K). For this, a more exhaustive analysis will be needed.

ISBN 978-84-947240-2-2

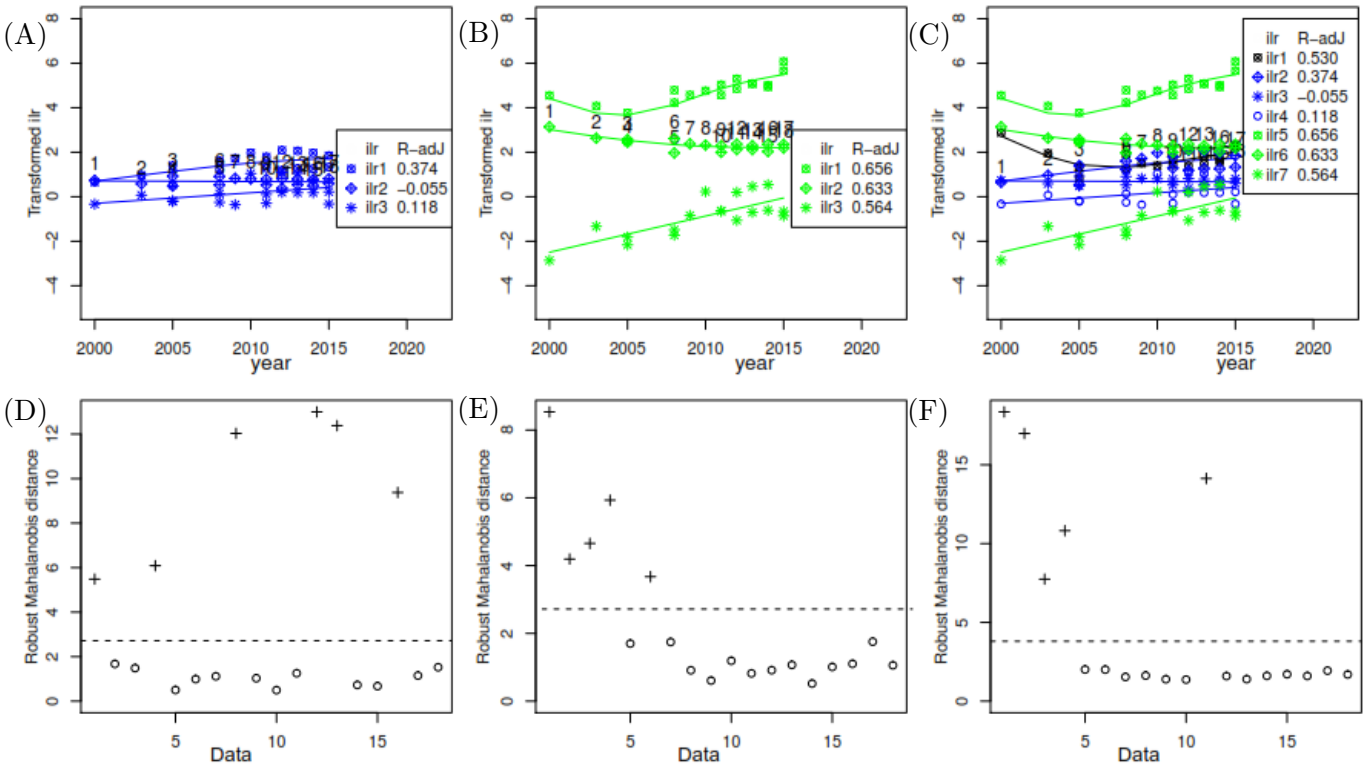


Figure 3: Outliers Colombia: (A), (D) Rural; (B), (E) Urban; (C), (F) Aggregate Urban/Rural.

The number of outliers detected has been variable. Figure 3D, 3E, and 3F show that this variation has not been punctual but in the time series. On the other hand, we expected less presence of outliers in aggregate data, due to the transforms; and a marked difference in the amount detected, between CoDa 4 and CoDa8, but this did not happen (See Table 1).

Table 1: Number of outliers detected in the countries.

Country/ N° Outlier	Rural (CoDa4)	Urban (CoDa 4)	Rural and Urban (CoDa 8)
Bolivia	4	**0**	5
Colombia	6	**5**	6
Ecuador	5	**4**	**4**
Paraguay	6	5	**4**
Peru	**5**	8	7
Uruguay	**1**	4	ND

A vertical analysis of table 1 shows that in the rural sector there is a greater presence of outliers in Colombia and Paraguay; being this six. While Uruguay presents a single outlier. In the urban sector, a greater number of outliers have been detected in Peru, while Bolivia has no value.

he horizontal analysis of table 1 shows the low quantity of outliers in disaggregated data (Urban and Rural) of Bolivia, then in aggregate data (CoDa 8). In Colombia, less is seen in the urban sector. In Ecuador, it could be assessed with less presence in aggregate data. In Paraguay there is less in aggregate data.

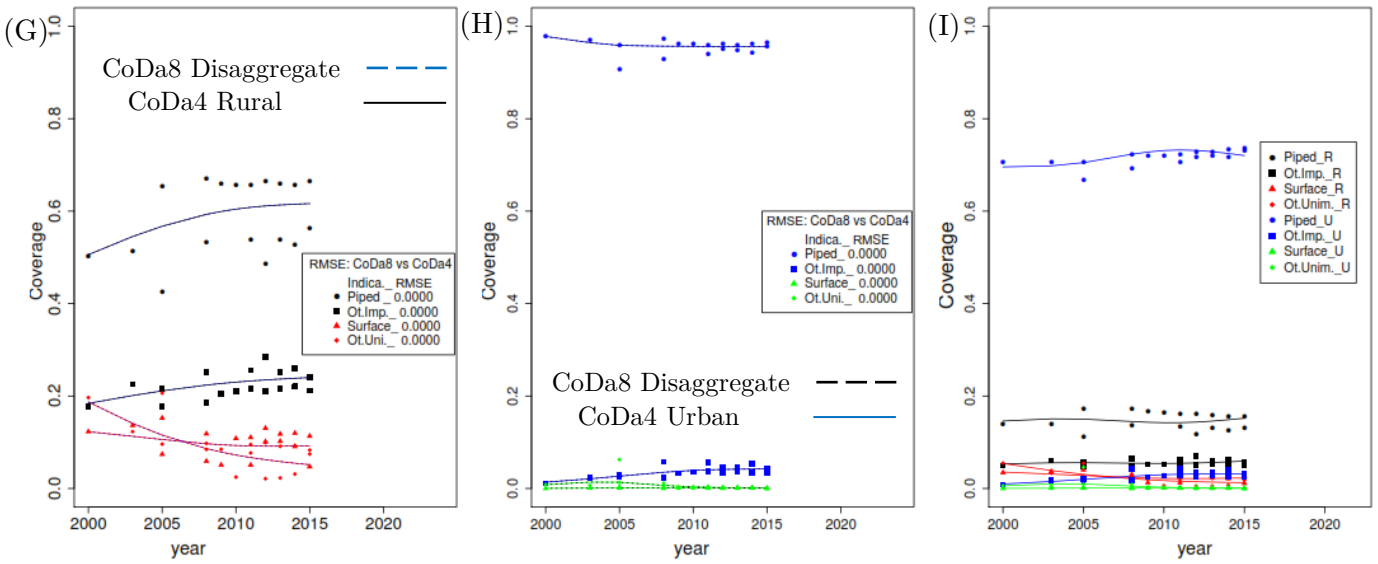
In summary, there has been no significant difference in the amount of data with outliers that allows us to assess whether four-part CoDa or eight-part CoDa is better.

ISBN 978-84-947240-2-2

3.2 Aggregated data vs. disaggregated data

The aggregate data analysis (Figure 4I) for scenario 1 shows the predominance in the proportion of the urban population that access piped water services. In addition, it can be seen that the proportion of people with access to unimproved water is low compared to the total.

In Figure 4G and 4H, it can be seen that the estimated values in CoDa 4 and CoDa 8 show the same trends in all indicators (solid line equal to the dashed line). The same analysis was carried out for the countries under evaluation, observing that in all of them the RMSE is equal to zero (See table 2). Which leads us to conclude that under the partition performed (Figure V1, V2 and V3) in scenario 1, it does not matter if we do the analysis in CoDa of four or eight parts, the result will always be the same.



(J) (K) (L)

CoDa8 Disaggregate
CoDa4 Rural

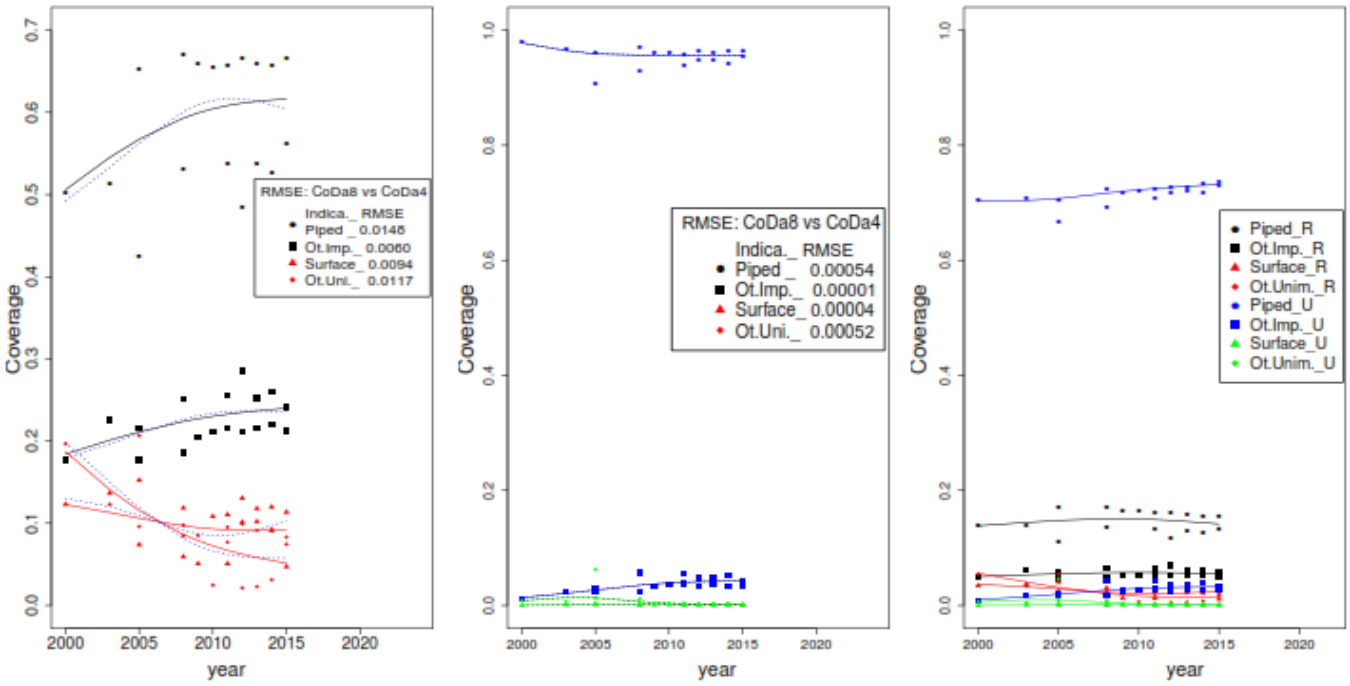


Figure 4: Comparison of models in aggregate and disaggregated data - Colombia. **Scenario 1**: (G) Rural disaggregate; (H) Urban disaggregate; (I) Rural and Urban aggregate. **Scenario 2**: (J) Rural disaggregate; (K) Urban disaggregate; (L) Rural and Urban aggregate.

Opposite case it happens when the partition changes (Scenario 2, V4). This shows that there is variation when making estimates with aggregate data (Figure 4L) and individual estimates of four parts (Figure 4J and 4K). The comparison of both gives us RMSE values different from one (Figure 4J and 4K). This does not happen in all countries or in all indicators (See table 2). In this exception is Paraguay, in which the RMSE remains zero in both scenarios (See table 2).

Table 2: Comparative of estimates between four-part CoDa and eight-part CoDa

País/(RMSE x 10^{-2})		Xr_1	Xr_2	Xr_3	Xr_4	Xu_1	Xu_2	Xu_3	Xu_4
Bolivia	Esc. 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Esc. 2	0.2761	0.1381	0.285	0.1214	0.0507	0.0027	0.0065	0.0471
Colombia	Esc. 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Esc. 2	1.4868	0.6074	0.9495	1.1742	0.0549	0.0013	0.0045	0.0524
Ecuador	Esc. 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Esc. 2	0.5997	0.1628	0.4873	0.2751	0.154	0.0092	0.016	0.1479
Paraguay	Esc. 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Esc. 2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Peru	Esc. 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Esc. 2	0.4504	0.1625	0.3023	0.3054	0.1416	0.0098	0.0135	0.1379

ISBN 978-84-947240-2-2

Uruguay	Esc. 1	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Esc. 2	0.7032	0.1292	0.089	0.7448	0.0479	0.0002	0.0017	0.0464

In Paraguay the change of SBP has had no effect on the transformations; consequently, the estimates in CoDa 4 are the same as when disaggregated estimates of CoDa eight.

The variation between the two (aggregate, disaggregated data) has not been significant either, due to the fact that most of the subcompositions have RMSE values close to zero. After this analysis, we compare the efficiency of the prediction models in disaggregated data with the results of estimates in scenario 2.

Table 3: Comparison of the best efficiency of the models in aggregate and disaggregated data (Scenario 2).

País/NSE		**Xr_1**	**Xr_2**	**Xr_3**	**Xr_4**	**Xu_1**	**Xu_2**	**Xu_3**	**Xu_4**
Bolivia	CoDa4	-0.014	**0.508**	**0.375**	0.274	0.207	0.589	0.35	0.639
	CoDa8	**-0.008**	0.503	0.361	**0.278**	**0.215**	0.589	**0.358**	**0.653**
Colombia	CoDa4	0.15	**0.252**	**0.089**	0.512	**0.087**	0.49	**0.078**	**0.281**
	CoDa8	**0.169**	0.238	0.083	**0.528**	0.078	0.49	0.074	0.269
Ecuador	CoDa4	**0.453**	0.335	**0.607**	**0.476**	0.35	**0.287**	0.016	-0.008
	CoDa8	0.421	**0.343**	0.597	0.427	**0.392**	0.284	**0.089**	**0.002**
Paraguay	CoDa4	0.915	0.535	0.816	0.918	0.653	0.573	0.158	0.52
	CoDa8	0.915	0.535	0.816	0.918	0.653	0.573	0.158	0.52
Peru	CoDa4	**0.528**	0.398	**0.383**	**0.665**	-0.001	-0.036	-0.023	**0.674**
	CoDa8	0.52	**0.408**	0.367	0.659	**0.005**	-0.036	**-0.011**	0.666
Uruguay	CoDa4	**0.609**	0.366	**0.55**	**0.576**	0.138	**0.92**	**-0.112**	0.027
	CoDa8	0.594	**0.373**	0.546	0.548	**0.165**	0.919	-0.135	**0.056**

Horizontal analysis of table 3 shows that the predictive capacity of CoDa 8 in Bolivia is slightly better than CoDa 4. NSE values are higher in five out of eight indicators. The opposite situation occurs in Colombia, in which CoDa 4 presents better predictive capacity than CoDa 8. In Ecuador, both show improvements in four out of eight indicators. Therefore, it would have to be analyzed with other metrics that help to clarify whether CoDa 8 or CoDa 4 is better. Paraguay, is not affected by the variation in the partitions. In addition, it gives the same result if CoDa 8 or CoDa 4 (Table 2) is used. In Peru, CoDa 4 has a better predictive capacity. In Uruguay, CoDa 4 is better than CoDa 8.

In summary, the comparison of analyzing data in aggregate or disaggregated form, indicates that by doing analysis with CoDa 4, better predictive capacity will be presented in Uruguay, Peru, and Colombia.

For Paraguay, the use of either of them is indifferent. In Ecuador a more exhaustive analysis must be done, a priori NSE values are better in four out of eight indicators in rural and urban areas. Bolivia is the only country in which the eight-part CoDa analysis is better because its predictive capacity is better in most indicators.

ISBN 978-84-947240-2-2

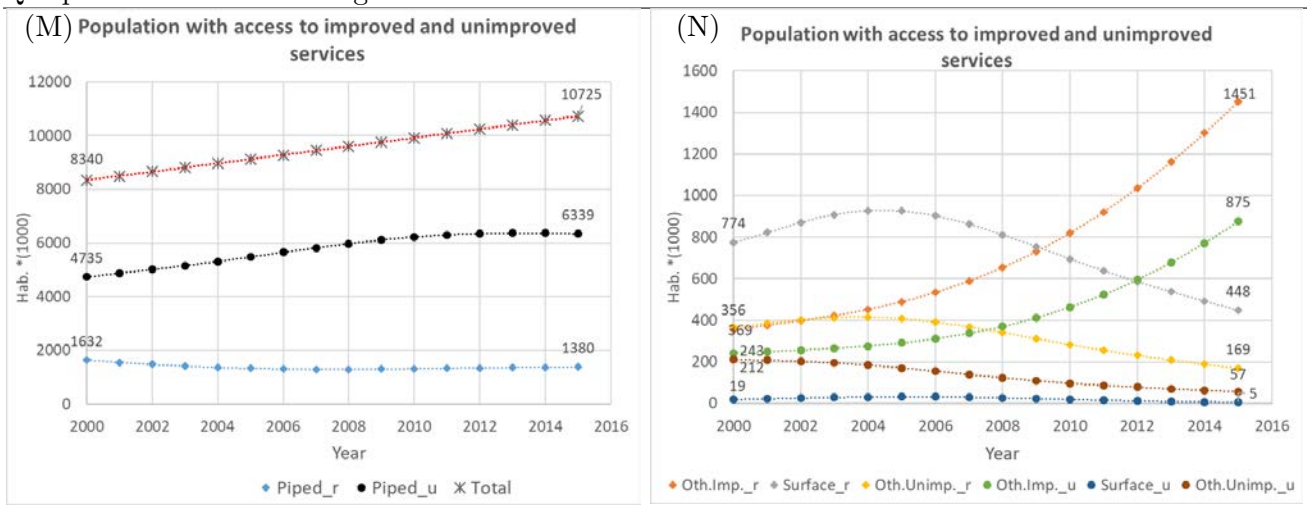


Figure 5: Temporal evolution in aggregate data of the population that accesses water services in Bolivia.

The temporal evolution of the population that accesses different water sources in Bolivia are shown in Figure 5 (Results of analysis in CoDa 8). The urban population that accesses piped water is increasing (Figure 5M). While the rural population in the same indicator shows a slight decrease. This decrease is offset by the increase in the population accessing water through other improved forms (Figure 5N). Of these, there was a greater increase in the rural population, reaching 1.451 million people who access this service. On the other hand, the urban and rural population that accesses surface water shows significant decreases in recent periods. Despite this, there is a large number of people from rural areas who continue to drink from these sources (448 thousand – Year 2015). The population that accesses water from unimproved sources decreased in the urban sector to 57 thousand people; while in the rural sector 169 thousand people were reached. Despite these decreases in both indicators (access to water by other improved forms, and access to surface water), in the rural sector more efforts are needed to close the existing gaps. Moreover, it is the area in which there is a high poverty rate and in the new SDGs they are focused on "nobody being left behind" (ONU, 2015).

In summary, the analysis in aggregate data has allowed us to analyze the temporal evolution of the population that accesses different water sources. In addition, with this methodology of work in the sector, national estimates are simple to perform, because it is the simple sum of the subcompositions; which does not happen at the moment (JMP, 2018).

Estimated values in this document may differ from the international JMP report, because linear regression methods are used, whereas GAM is used in this study. This has already been discussed extensively in the literature (Fuller et al., 2016; Bartram et al., 2014; Wolf et al., 2013; Pérez-Foguet et al., 2017).

4 Conclusions

It has been shown that in scenario 1 with an SBP1, using CoDa 4 or CoDa 8, the estimates give the same results (Figure 4G and 4H). While in scenario 2 with an SBP2, they present small differences (Figure 4J and 4K). This leads us to conclude that the selection of the PBS will influence the estimates. As a result, it opens a lot of possibilities to do analysis with different SBP and make the selection that best predictive capacity present. It is suggested to do these tests only in case you do not look for interpretations in the transforms. Otherwise, the appropriate group of parts should be selected to help us interpret better in the transforms. Because these carry proportions that contain information.

ISBN 978-84-947240-2-2

Evidence shows that CoDa 4 usually fares better in Uruguay, Peru, and Colombia (Table 3). While, in Bolivia, CoDa 8 presents better predictive capacity in five out of eight indicators. In Ecuador, you cannot infer, which of them is better.

The aggregate analysis in the data (CoDa 8), has allowed us to know in full, the temporal evolution of the population that accesses different water sources. A particular case is the one addressed in Bolivia. In which, there is an increase in the rural and urban population that accesses water by other improved forms. This was compensated by the decrease in access to unimproved water. On the other hand, it was found mostly in the rural sector, populations that access surface water sources. Consequently, Bolivia's agenda should be aimed at closing gaps in water, sanitation and hygiene. Taking as criteria, the poorest and most vulnerable populations.

On the other hand, the use of CoDa in aggregate data has certain disadvantages. The main one is the loss of information because it cannot complete the composition if it lacks data in some variables of the total. In the disaggregated analysis (urban and rural), the possibility of affecting only one sector is presented; what leads to not losing information in the other and consequently do the common analysis.

Regarding outliers, it cannot be inferred whether CoDa 8 or CoDa 4 has a lower quantity (Table 1) because there has not been significant variation. In later studies, comparisons of the models will be made by removing the outliers in each scenario.

Acknowledgements and appendices

The programming script is posted in GitHub (https://github.com/fquispec/Congress_CoDa_2019)

References

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

Bartram, Jamie, Clarissa Brocklehurst, Michael Fisher, Rolf Luyendijk, Rifat Hossain, Tessa Wardlaw, and Bruce Gordon. (2014). "Global Monitoring of Water Supply and Sanitation: History, Methods and Future Challenges." *International Journal of Environmental Research and Public Health* 11 (8): 8137–65. https://doi.org/10.3390/ijerph110808137.

Boogaart, K. Gerald van den, Raimon Tolosana-Delgado, and Matevz Bren. (2014). "Compositions: Compositional Data Analysis." Boston. https://cran.r-project.org/package=compositions.

Cohen, Barney. (2004). "Urban Growth in Developing Countries: A Review of Current Trends and a Caution Regarding Existing Forecasts." *World Development* 32 (1): 23–51. https://doi.org/10.1016/J.WORLDDEV.2003.04.008.

Dirven, Martine, Rafael Echeverri, Cristina Sabalain, David Candia Baeza, Sergio Faiguenbaum, Adrián G. Rodríguez, and Carolina Peña. (2011). "Hacia Una Nueva Definición de 'Rural' Con Fines Estadísticos En América Latina." CEPAL. https://repositorio.cepal.org/handle/11362/3858.

DS. N°031-2008-VIVIENDA. (2008). "Decreto Supremo Que Modifica El Texto Único Ordenado Del Reglamento de La Ley General de Servicios de Saneamiento." Lima. http://www.sedapal.com.pe/contenido/031-2008-VDA (30.11.2008).pdf.

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. https://doi.org/10.1023/A:1023818214614.

Egozcue, Juan Jose, and Vera Pawlowsky-Glahn. (2005). "Groups of Parts and Their Balances in Compositional Data Analysis." *Mathematical Geology* 37 (7): 795–828. https://doi.org/10.1007/s11004-005-7381-9.

Filzmoser, Peter, and Karel Hron. (2008). "Outlier Detection for Compositional Data Using Robust Methods." *Mathematical Geosciences* 40 (3): 233–48. https://doi.org/10.1007/s11004-007-9141-5.

Fuller, James A., Jason Goldstick, Jamie Bartram, and Joseph N.S. Eisenberg. (2016). "Tracking Progress towards Global Drinking Water and Sanitation Targets: A within and among Country Analysis." *Science of The Total Environment* 541 (January): 857–64. https://doi.org/10.1016/J.SCITOTENV.2015.09.130.

INE. (2018). "MEMORIA CENSO 2017 - GLOSARIO." http://www.censo2017.cl/memoria/descargas/memoria/libro_glosario_censal_2017.pdf.

JMP. (2017). "Joint Monitoring Programme for Water Supply, Sanitation and Hygiene." 2017. https://washdata.org/data.

JMP. (2018). "JMP Methodology 2017 Update & Sdg Baselines." https://washdata.org/sites/default/files/documents/reports/2018-04/JMP-2017-update-methodology.pdf.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn. (2003). "Dealing with Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation." *Mathematical Geology* 35 (3): 253–78. https://doi.org/10.1023/A:1023866030544.

Martín-Fernández, Josep Antoni, Javier Palarea-Albaladejo, and Ricardo Antonio Olea. (2011). "Dealing with Zeros." In *Compositional Data Analysis*, 43–58. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119976462.ch4.

ONU. (2015). "Transformar Nuestro Mundo: La Agenda 2030 Para El Desarrollo Sostenible." *Asamblea General. Septuagésimo Período de Sesiones de La Asamblea General de Las Naciones Unidas, Del 11 Al 18 de Septiembre Del 2015 (Resolución A/RES/70/1)* 16301: 40. https://doi.org/http://unctad.org/meetings/es/SessionalDocuments/ares70d1_es.pdf.

Palarea-Albaladejo, J., and J.A. Martín-Fernández. (2008). "A Modified EM Alr-Algorithm for Replacing Rounded Zeros in Compositional Data Sets." *Computers & Geosciences* 34 (8): 902–17. https://doi.org/10.1016/J.CAGEO.2007.09.015.

Pérez-Foguet, A., R. Giné-Garriga, and M.I. I. Ortego. (2017). "Compositional Data for Global Monitoring: The Case of Drinking Water and Sanitation." *Science of the Total Environment* 590–591 (July): 554–65. https://doi.org/10.1016/j.scitotenv.2017.02.220.

Pinheiro, Jose, Douglas Bates, Saikat DebRoy, Sarkar Deepayan, and {R Core Team}. (2018). "Linear and Nonlinear Mixed Effects Models." https://cran.r-project.org/package=nlme.

R Core Team. (2018). "R: A Language and Environment for Statistical Computing." Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org/.

Templ, Matthias, Karel Hron, and Peter Filzmoser. (2011). "RobCompositions: An R-Package for Robust Statistical Analysis of Compositional Data." In *Compositional Data Analysis: Theory and Applications*, 341–55. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119976462.ch25.

Templ, Matthias, Karel Hron, Peter Filzmoser, and Alžběta Gardlo. (2016). "Imputation of Rounded Zeros for High-Dimensional Compositional Data." *Chemometrics and Intelligent Laboratory Systems* 155 (July): 183–90. https://doi.org/10.1016/J.CHEMOLAB.2016.04.011.

WHO/UNICEF. (2015). "Progress on Sanitation and Drinking Water – 2015 Update and MDG Assessment." World Health Organization and UNICEF. 2015. https://washdata.org/reports.

WHO/UNICEF. (2017). "Progress on Drinking Water, Sanitation and Hygiene – 2017 Update and SDG Baselines." World Health Organization and UNICEF. 2017. https://washdata.org/reports.

Wolf, Jennyfer, Sophie Bonjour, and Annette Prüss-Ustün. (2013). "An Exploration of Multilevel Modeling for Estimating Access to Drinking-Water and Sanitation." *Journal of Water and Health* 11 (1): 64–77. https://doi.org/10.2166/wh.2012.107.

Wood, Simon N. (2017). "Generalized Additive Models: An Introduction with R (2nd Edition)." *Chapman and Hall/CRC*.

ISBN 978-84-947240-2-2

Assessing CoDa regression for modelling daily multivariate air pollutants evolution

Joseph Sánchez-Balseca[1], **Agustí Pérez-Foguet**[1]
[1]Engineering Sciences and Global Development (EscGD),
Civil and Environmental Engineering Department,
Universitat Politècnica de Catalunya, Barcelona, Spain.
joseph.sanchez@upc.edu

Summary

The application of the theory of compositional data in multivariate spatio-temporal statistical models is still scarce, even though the results obtained are robust. Actually, this kind of models are attractive to pollution model developers, due to, its versatility in the spatio-temporal variables; but nobody has tried to use it with compositional data yet. The main differences between a conventional model and two CoDa models (with two sequential binary partition, SBP) were analyzed. The first SBP was proposed by pollutants relationship interpretation, and the second one was imposed as standard SBP (R studio). Initially the conventional temporal model is used to predicting pollution levels to fill missing data or predicting pollution levels on future days. The application of compositional data theory in conventional temporal air quality models allowed to obtain acceptable quality models, whose results were adjusted to the observed values. Nash-Sutcliffe Efficiency Index (NSE) and root-mean-square error (RMSE), were used to evaluating the model quality and fitted values respectively.

Key words: air quality, compositional data, SBP, multivariate response model, precision.

1 Introduction

Predictions from numerical models are also used for environmental regulatory purposes and improved decision-making strategies (Zannetti, 1990; Mayer 1999; Dominici et al. 2002; Cetin et al. 2017; Paci, 2013). Shaddick G. et al. (2002) studied the hierarchical Dynamic Linear Model (DLM) applied to four pollutants, at eight monitor stations. Gutierrez et al. (2016), proposed to model the measurements of particulate matter, by means of a Bayesian nonparametric dynamic model. Recently, Shaddick et al. (2018), developed a spatially-varying model, within a Bayesian hierarchical modelling framework.

Most data in the geo-environmental sciences are compositional in character. They describe quantitatively the parts of a whole. If concentrations are not considered as compositional data, incorrect conclusions could be obtained (Egozcue and

ISBN 978-84-947240-2-2

Pawlowsky-Glahn, 2011). Due to the advances in the compositional data analysis since 2000, the statistical works can be resolved in three steps: data transformations to log-ratio coordinates, traditional statistical analysis with the coordinates, and finally result analysis (Aitchison et al., 2002; Egozcue et al., 2003; van den Boogaart and Tolosana, 2013).

The present work has been structured as: Methodology (section 2), Results (section 3), and Conclusions (section 4).

2 Methodology

The methodology will begin with a data description, then it will explain the hierarchical Dynamic Linear Model theory, the compositional data concepts used, and finally a numerical comparison analysis (models).

2.1 Data

The data used in the temporal multivariate linear models was collected hourly over the period 2009-2013, from three air quality monitoring stations at Quito-Ecuador (Environmental Department of Quito, 2017). The main stations characteristics are showed in the table 1.

Table 1: Main parameter of three monitor stations.

Station Name	Location	High (mals)	Station code
CARAPUNGO	78°26'50'' W, 0°5'54'' S	2660	1
BELISARIO	78°29'24'' W, 0°10'48'' S	2835	2
EL CAMAL	78°30'36'' W, 0°15'00'' S	2840	3

This tree station covering the urban Quito territory, and are located in the north (CARAPUNGO/Station Nº 1), in the center (BELISARIO/Station Nº 2), and south (EL CAMAL/Station Nº 3); as it is showed in the figure 1.

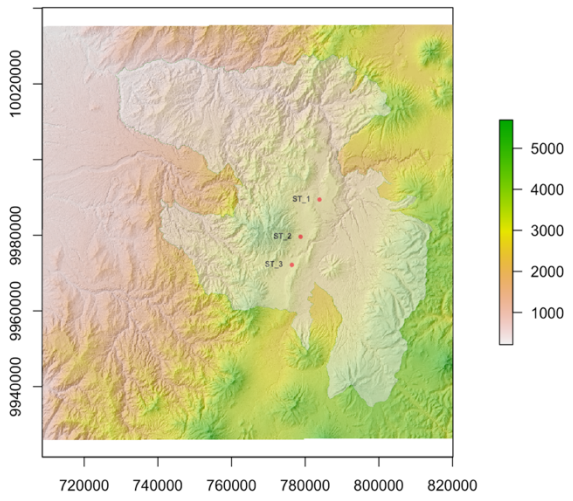


Figure 1: Monitoring station location

ISBN 978-84-947240-2-2

The data analyzed as dependent variable were carbon monoxide (CO), nitrogen dioxide (NO_2), Sulphur dioxide (SO_2), and ozone (O_3). To obtain the daily mean concentration, the presence of 75% hourly data, was imposed. The variable with more missing daily values was nitrogen dioxide (2.5%, 1.1%), at station 1 and 2 respectively. The data description is showed in the table 2, where the units to pollutants are given as parts per billion (ppb), and Kelvin units (K) for temperature.

Table 2: Summary of variables measured at three monitor stations, 2009-2013.

Station Nº 01									
Var.	Total	Missing	%	Mean	Min.	25%	Median	75%	Max.
CO	1827	37	2.0	640	106.6	501.1	616.6	749.5	1940.6
NO_2	1827	45	2.5	11.8	0.5	8.6	11.2	14.3	44.9
SO_2	1827	27	1.5	1.6	0.05	0.95	1.5	2.1	6.6
O_3	1827	24	1.3	18.2	0.0	14.8	17.4	20.6	46.3
T	1827	5	0.3	14.5	11.3	13.8	14.5	15.1	18.4
Station Nº 02									
Var.	Total	Missing	%	Mean	Min.	25%	Median	75%	Max.
CO	1827	15	0.8	955.7	250.2	749.4	921.4	1111.8	2497.2
NO_2	1827	21	1.1	19.83	5.362	16.32	19.498	23.024	74.880
SO_2	1827	17	0.9	2.604	0.119	1.673	2.468	3.282	10.462
O_3	1827	14	0.7	14.51	0.389	10	13.246	17.057	56.399
T	1827	5	0.3	14	10.19	13.22	14.06	14.83	17.92
Station Nº03									
Var.	Total	Missing	%	Mean	Min.	25%	Median	75%	Max.
CO	1827	15	0.8	947.3	157.9	775.1	929.8	1100.4	2254.4
NO_2	1827	17	0.9	21.73	6.952	18.28	21.398	25.073	42.299
SO_2	1827	19	1.0	4.162	0.296	2.124	3.1486	5.0276	28.6855
O_3	1827	14	0.7	16.03	4.25	11.80	14.87	18.36	57.98
T	1827	6	0.3	14	10.69	13.46	14.19	14.89	17.60

2.2 Dynamic Linear Models/Temporal air quality model

The most widely known applied subclass is that of normal dynamic linear models, referred as dynamic linear models, or DLMs (West and Harrison, 1997). For this work, Y_{pt} represent the pollutant p concentration, on day t, the observation [Eq. (3)] and system [Eq. (4)] equations respectively are

$$Y_{pt} = \theta_{pt} + v_{pt} \qquad v_{pt} \sim N(0, \sigma_{vp}^2) \tag{3}$$

$$\theta_{pt} = \theta_{p,t-1} + w_{pt} \qquad w_{pt} \sim N(0, \sigma_{wp}^2) \tag{4}$$

The conditional variance σ_w^2 is not comparable with σ_v^2. The multiple pollutants at any time point were modelled as arising from a multivariate Gaussian random field. v_t represents the measurement errors which are assumed to be independent and identically distributed $N(0, \sigma_{vp}^2)$; w_t to each pollutant are independent and identically distributed multivariate normal random variables with zero mean and variance-covariance matrix Σ_p. To this work a Bayesian approach was adopted,

ISBN 978-84-947240-2-2

with prior distribution being assigned to the unknown parameters and missing observations. Gamma priors are proposed for the precisions $\sigma_v^{-2} \sim Ga(a_v, b_v)$, in the multivariate updating scenario, the variance-covariance matrix $\Sigma_p^{-1} \sim W_p(D, d)$, where $W_p(D, d)$ denotes a *P*-dimensional Wishart distribution with mean *D* and precision parameter *d*.

2.3 Compositional Data

In mathematical terms, compositional data are represented [Eq. (6)] as pertaining to a sample space called the simplex S^D

$$S^D = \{x = (x_1, x_2, x_D): x_i > 0 (i = 1,2,D), \sum_{i=1}^{D} X_i = K\} \quad (6)$$

where *K* is a given positive constant, defined a priori and depending on how the parts are measured (Buccianti, 2013). This work has four gaseous pollutants (ppb), and the composition to be analysed would be $[CO, NO_2, SO_2, O_3, R]$, in this case the residual part was not of interest, for this reason the subcompositional approach is used (the subcomposition closure for each day is saved to be used in the model evaluation). The subcompositional incoherence, was eliminated through log-ratio methods (Buccianti et al., 2006). Aitchison (1982) developed the additive–log-ratio (alr) and centred-log-ratio (clr) transformations; Egozcue et al. (2003) introduced the isometric-log-ratio (ilr) transformation.

The isometric-log ratio (ilr) transformation, was used. In this framework, the procedure of the sequential binary partition (SBP) to identify orthonormal coordinates was adopted (Egozcue and Pawlowsky-Glahn, 2005). For this study, the standard base (function in R-studio) and SBP method [Eq. (6)] were used. Considering the four pollutants as sub composition (XCO, XNO_2, XSO_2, XO_3), in first level, the group was divided as: SO_2, O_3 and CO, NO_2. One group is related by patterns that showed air pollutant pairs of O_3/SO_2 appearing at the same hour of the day (Meagher et al., 1827; EPA, 2001).

$$x_1^* = \sqrt{\frac{2 \cdot 2}{2+2}} ln \frac{(XCO \cdot XNO_2)^{\frac{1}{2}}}{(XSO_2 \cdot XO_3)^{\frac{1}{2}}}$$

$$x_2^* = \sqrt{\frac{1 \cdot 1}{1+1}} ln \frac{XCO}{XNO_2} \quad (7)$$

$$x_1^* = \sqrt{\frac{1 \cdot 1}{1+1}} ln \frac{XSO_2}{XO_3}$$

2.4 Comparison Models

To evaluating and comparing the models, Nash-Sutcliffe Efficiency Index (NSE) and root-mean-square error (RMSE), were used. NSE is a widely used and potentially reliable statistic for assessing the goodness of fit of hydrologic models.

ISBN 978-84-947240-2-2

The NSE scale is from 0 to 1 [Eq. (8)], NSE = 1 means the model is perfect. NSE = 0 means that the model is equal to the average of the observed data, and negative values mean that the average is a better predictor (McCuen et al., 2006).

$$NSE = 1 - \frac{\sum(Yobs_i - Ysim_i)^2}{\sum(Yobs_i - \overline{Yobs})^2} \qquad (8)$$

Root Mean Square Error (RMSE) is a measure of quantitative performance commonly used to evaluate demand forecasting methods. The lower RMSE, it means that the model has no error [Eq. (9)].

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Ysim_i - Yobs_i)^2}{n}} \qquad (9)$$

3 Results

The conventional model for nitrogen dioxide presented adequate fitted values; in the station 1, it presented the less variability value, and its behavior for a subset of recorded measurements (90% confidence interval band) and estimated θ for station 1 is showed in the figure 2.

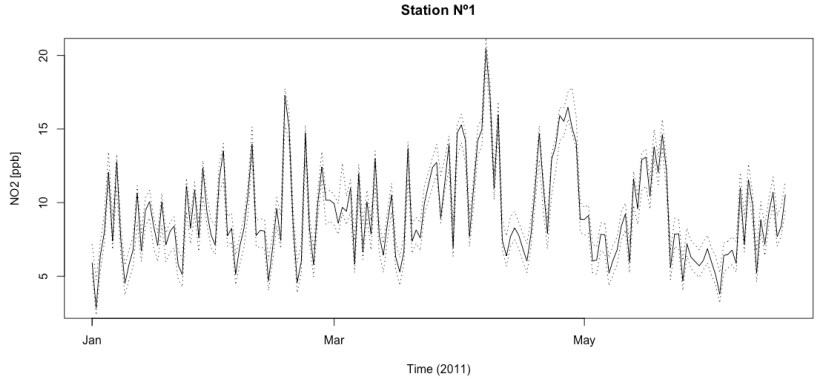


Figure 2: Conventional temporal model (year: 2011; 90% confidence interval)

Threshold value suggested to indicate a model of sufficient quality is NSE>0.5 (Ritter and Muñoz-Carpena, 2013). In general, the conventional model had low NSE values. If threshold value is used, the models to evaluating the behavior of CO at station 1 (NSE=0.43), and SO2 at station 3 (NSE=0.45) have not sufficient quality. In comparison with the compositional models, which had NSE values over the threshold for all pollutants (NSE≥0.62).

In general, a lower RMSE is better than a higher one. If RMSE is zero, the model has perfect fit to the data. The RMSE values for conventional model were high, except to SO2 at stations 1 and 2. The compositional model (RMSE ≤ 0.009) presented a better fitted data for all pollutants. For this work, two SBP were used, therefore similar results were obtained, the littles differences found among them are showed in the figure 3, where the residuals of CO and NO_2 had higher values than SO_2 and O_3. These differences were generated in the error treatment over transformations.

ISBN 978-84-947240-2-2

The two compositional models presented acceptable fitted values, more than conventional model. The NSE and RMSE mean values are showed in the table 8.

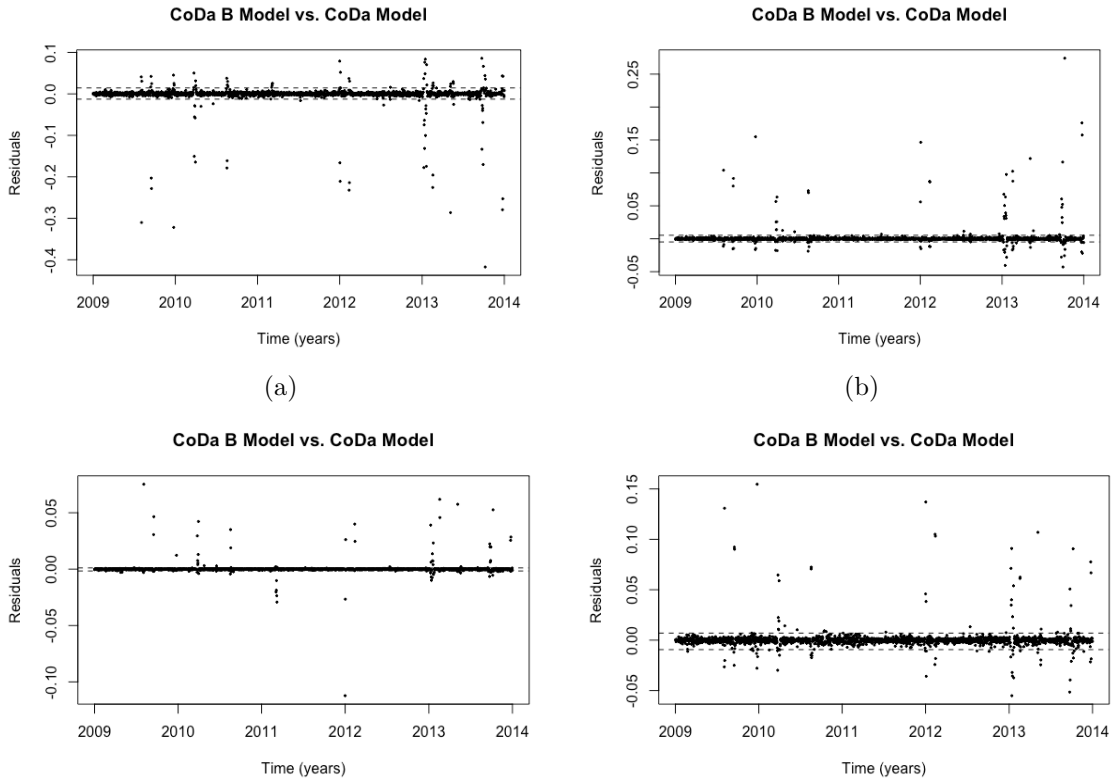


Figure 3: Residuals between compositional models for station 1:
(a) CO, (b) NO_2, (c) SO_2, and (d) O_3

Table 8: NSE and RMSE mean values for each model.

	Model		Compositional model (standard base)		Compositional model (base imposed)	
STATION	NSE	RMSE	NSE	RMSE	NSE	RMSE
1	0,6867009	39,04213233	0,759960075	0,004501534	0,929408425	0,002467467
2	0,79693	33,47364475	0,924419225	0,002468459	0,952827725	0,001662221
3	0,681427375	3,118657	0,981204225	0,001408042	0,84546585	0,004259196
MEAN	0,721686092	25,21147803	0,888527842	0,002792679	0,909234	0,002796295

4 Conclusions

The compositional data concepts applied to temporal models (Dynamic Linear Models) presented good results, like as better-quality models and better fitted values. These results are due to use the air pollutants concentrations as compositions. Compositional models had less variability than conventional models, over all pollutants at three stations.

ISBN 978-84-947240-2-2

Acknowledgements and appendices

The authors would like to thank the "Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación de la República del Ecuador" for financial support.

References

Aitchison, J. 1982. «The statistical analysis of compositional data (with discussion)». *Journal of the Royal Statistical Society* 139-177.

Aitchison, J., Barceló-Vidal, C., J.J. Egozcue, y V. Pawlowsky- Glahn. 2002. «A concise guide for the algebraic-geomet- ric structure of the simplex, the sample space for compo- sitional data analysis». *The eighth annual con- ference of the international Association for Mathematical Geology, Volume I and II.* Berlín: Selbstverlag der Alfred-Wegener-Stiftung. 387-392.

Akpinar, Ebru, Sinan Akpinar, y Hakan Öztop. 2009. «Statistical analysis of meteorological factors and air pollution at winter months in Elaziğ, Turkey.» *Journal of Urban and Environmental* 7-16.

Buccianti, A., G. Mateu, y V. Pawlowsky. 2006. *Compositional Data Analysis in the Geosciences.* London: Geological Society.

Buccianti, Antonella. 2013. «Is compositional data analysis a way to see beyond the illusion?» *Computers y Geosciences* 165-173.

Dominici, Francesa, Aidan McDermott, Scott Zeger, y Jonathan Samet. 2002. «On the Use of Generalized Additive Models in Time-Series Studies of Air Pollution and Health.» *American Journal of Epidemiology* 193–203.

Egozcue, J. J., y V. Pawlowsky-Glahn. 2011. «Análisis composicional de datos en Ciencias Geoambientales.» *Boletín Geológico y Minero* 439-452.

Egozcue, J., y V. Pawlowsky-Glahn. 2005. «Groups of parts and their balances in compositional data analysis.» *Mathematical Geology* 795-828.

Egozcue, J.J., V. Pawlowsky-Glahn, G. Mateu-Figueras, y C. Barceló-Vidal. 2003. «Isometric logratio transformations for compositional data analysis.» *Mathematical Geology* 279-300.

Egozcue, Juan José, Josep Daunis-i-Estadella, Vera Pawlowsky-Glahn, Karel Hron, y Peter Filzmoser. 2012. «Simplicial regression. The normal model.» *Journal of Applied Probability and Statistics* 87-108.

Environmental Protection Agency. 2001. *AQS* . Report, Washington: EPA.

Gutiérrez, Luis, Ramsés H. Mena, y Matteo Ruggiero. 2016. «A time dependent Bayesian nonparametric model for air quality analysis.» *Computational Statistics and Data Analysis* 161–175.

Marinov, Marin, Ivan Tapalov, Elitsa Gieva, y Georgi. Nicolov. 2016. «Air quality monitoring in urban environments.» 2016 39th International Spring Seminar on Electronics Technology (ISSE). Pilsen: IEEE. 443 - 448.

Mayer, Helmut. 1999. «Air pollution in cities.» *Atmospheric Environment* 4029-4037.

McCuen, Richard H., Zachary Knight, y A. Gillian Cutter. 2006. «Evaluation of the Nash–Sutcliffe Efficiency Index.» *Journal of Hydrologic Engineering.*

Meagher, J. F., N. T. Lee, R. J. Valente, y W. J. Parkhurst. 1967. «Rural ozone in

ISBN 978-84-947240-2-2

the southeastern United States. .» *Atmospheric Environment* 605–615.

Paci, Lucia. 2013. *Bayesian space-time data fusion for real-time forecasting and map uncertainty.* Bologna: Università di Bologna.

Ritter, Axel, y Rafael Muñoz-Carpena. 2013. «Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments.» *Journal of Hydrology* 33-45.

Sarstedt, M., y E. Mooi. 2014. «A Concise Guide to Market Research.» *Business and Economics* 273-324.

Secretaria de Medio Ambiente DMQ. 2017. *Informe de Calidad del Aire DMQ 2017.* Quito: Distrito Metropolitano de Quito.

Shaddick, Gavin, Matthew L. Thomas, Amelia Jobling, Michael Brauer, Aaron van Donkelaar, Rick Burnett, Howard Chang, y otros. 2018. «Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution.» *Royal Statistical Society* 231–253.

Shaddick, Gavin, y Jon Wakefield. 2002. «Modelling Daily Multivariate Pollutant Data at Multiple Sites.» *Journal of the Royal Statistical Society* (Wiley) 351-372.

Tolosana, R., y H. Eynatten. 2010. «Simplifying compositional multiple regression: Application to grain size controls on sediment geochemistry.» *Computers & Geosciences* 577-589.

van den Boogaart, K. Gerald, y Raimon Tolosana-Delgado. 2013. *Analyzing Compositional Data with R.* Berlin: Springer.

West, Mike, y Jeff Harrison. 1997. *Bayesian Forecasting and Dynamic Models.* New York: Springer-Verlag.

Zannetti, Paolo. 1990. *Air pollution Modeling* . Monrovia, California: Springer Science.

ISBN 978-84-947240-2-2

Wavelet regression for compositional data

A. Srakar[1], and T.R.L. Fry[2]
[1]Institute for Economic Research and University of Ljubljana, Ljubljana, Slovenia; *andrej_srakar@t-2.net*
[2]RMIT University, Melbourne, Australia

Summary

Regression for compositional data has so far been largely considered only from a parametric point of view. Recently, some work adapted non-parametric regression to nonEuclidean manifolds. For example, Di Marzio et al. (2013) pursue the circular case, and Di Marzio et al. (2014) the spherical one. In a recent article, Di Marzio, Panzera and Venieri (2015) extended this to nonparametric situations, introducing local constant and local linear smoothing for regression with compositional data. Also, Barientos et al. (2015) propose a Bayesian nonparametric procedure for density estimation for data in a d-dimensional simplex. In our analysis, we extend the work of Di Marzio, Panzera and Venieri to locally adaptive estimators, in particular discrete and continuous wavelets. We rely on the work of Dey and Wang (2004), modeling the priors on triangles by use of wavelets constructed specifically for triangles. We transfer their methodology of deriving father and motherwavelets using a sequential approach to orthogonalization to derive the motherwavelets. Our new estimator is derived for three cases: simplicial-real; simplicial-simplicial; and real-simplicial regression and is based on Bayesian approach using wavelet type priors. We present a detailed statistical elaboration and analysis, simulation results to compare the performance with some existing parametric estimators for compositional data regression, and an application to the results to two case studies from economics– inference for inequality indices and international trade.

Key words: compositional data, regression, nonparametric, wavelets, Bayesian, expectation propagation

1 Introduction – regression for compositional data

A *composition* is defined as a vector of D positive components $x = (x_1, x_2, \dots, x_D)$ summing up to a given constant κ

It is generally – although not universally - agreed that the appropriate sample space for compositional data is the standard simplex (also called the "unit simplex"). It is defined as

$$S^D = \left\{ x = [x_1, x_2, \dots, x_D] \middle| x_i > 0, i = 1,2, \dots, D; \sum_{i=1}^{D} x_i = \kappa \right\}$$

For any vector of D real positive components

$$\boldsymbol{z} = [z_1, z_2, \dots, z_D] \in \mathbb{R}_+^D$$

($z_i > 0$ for all $i = 1,2, \dots, D$), the closure of $\boldsymbol{z}$ is defined as

$$\mathcal{C}(\boldsymbol{z}) = \left[\frac{\kappa \cdot z_1}{\sum_{i=1}^{D} z_i}, \frac{\kappa \cdot z_2}{\sum_{i=1}^{D} z_i}, \dots, \frac{\kappa \cdot z_D}{\sum_{i=1}^{D} z_i} \right]$$

Perturbation of a composition $x \in S^D$ by a composition $y \in S^D$ is defined as

$$x \oplus y = \mathcal{C}[x_1 y_1, x_2 y_2, \dots, x_D y_D]$$

ISBN 978-84-947240-2-2

Power transformation or powering of a composition $x \in S^D$ by a constant $\alpha \in \mathbb{R}$ is defined as

$$\alpha \odot x = \mathcal{C}[x_1^\alpha, x_2^\alpha, \dots, x_D^\alpha]$$

Regression with compositional data has been studied in multiple directions. Aitchison (2003) used classical methods on the log-ratio transformed space. Gueorguieva et al. (2008) applied Dirichlet regression. Stephens (1982) and Scealy and Welsh (2011) transformed the data on the surface of the unit hyper-sphere, using the square root transformation, and thus treated them as directional data. Tsagris (2015b) proposed the α-regression which relies upon the α-transformation (Tsagris et al., 2011). Compositional data regression from the Bayesian perspective was suggested by Shimizu et al. (2015). An important issue in compositional data is the presence of zeros, which cause problems for the logarithmic transformation. When zero values exist in data, Dirichlet models and the log-ratio transformation suggested by Aitchison (1982, 2003) and Egozcue et al. (2003) will not work unless a zero value imputation is applied first. Some important works on regression with compositional data including zeros, following imputation methods: Martin et al. (2003; 2012; 2018); Fry et al. (1996; 2000). As for the classification setting, Tsagris (2014) proposed the use of a power transformation applicable to cases with zero values in the data. Most papers focus on compositional data being in the response variable side. The case of compositional data in the predictor variables side was treated first by Hron et al. (2012) who applied the isometric log-ratio transformation to the compositional data and then applied a standard linear regression model. Di Marzio, Panzera and Venieri (2015) extended this to nonparametric situations, introducing local constant and local linear smoothing for regression with compositional data.

The problem of regression when the response is compositional is stated as follows. A compositional sample in S^D is available and it is denoted by $x_1, x_2, \dots, x_n$. The i-th data-point x_i is associated with one or more external variables or covariates grouped in the vector $\boldsymbol{t}_i = [t_{i0}, t_{i1}, \dots, t_{ir}]$, where $t_{i0} = 1$. The goal is to estimate the coefficients of a curve or surface in S^D whose equation is

$$\widehat{\boldsymbol{x}}(\boldsymbol{t}) = \beta_0 \oplus (t_1 \odot \beta_1) \oplus \cdots \oplus (t_r \odot \beta_r) = \oplus_{j=0}^{r} (t_j \odot \beta_j)$$

Deviation of this model from the data is defined as $\widehat{\boldsymbol{x}}(\boldsymbol{t_i}) \ominus x_i$ and its size by the Aitchison norm $\|\widehat{\boldsymbol{x}}(\boldsymbol{t_i}) \ominus x_i\|_a^2 = d_a^2(\widehat{\boldsymbol{x}}(\boldsymbol{t_i}), x_i)$. The target function (sum of squared errors, SSE) is

$$SSE = \sum_{i=1}^{n} \|\widehat{\boldsymbol{x}}(\boldsymbol{t_i}) \ominus x_i\|_a^2$$

to be minimized as a function of the compositional coefficients β_j which are implicit in $\widehat{\boldsymbol{x}}(\boldsymbol{t_i})$. The number of coefficients to be estimated in this linear model is $(r+1) \cdot (D-1)$.

Shimizu, Louzada, Suzuki, Ehlers (2015) derive a Bayesian analysis for compositional regression applying additive log-ratio (ALR) transformation and assuming uncorrelated and correlated errors.
As priors they use:

$$\beta_{0j} \sim N(a_{0j}, {b_{0j}}^2)$$
$$\beta_{lj} \sim N(a_{lj}, {b_{lj}}^2)$$
$${\sigma_j}^2 \sim IG(c_j, d_j)$$

Their joint posterior is:

$$\pi(\beta_{0j}, \beta_{lj}, {\sigma_j}^2|y) \propto \prod_{j=1}^{g} \exp\left[-\frac{1}{2{b_{0j}}^2}(\beta_{0j}-\alpha_{0j})^2\right] \times \prod_{j=1}^{g}\prod_{l=1}^{p} exp\left[-\frac{1}{2{b_{lj}}^2}(\beta_{lj}-\alpha_{lj})^2\right] \times \prod_{j=1}^{g}({\sigma_j}^2)^{-(c_j+1)}\exp(-d_j/{\sigma_j}^2) \times \prod_{j=1}^{g}({\sigma_j}^2)^{-\frac{n}{2}}\exp(-\frac{1}{2{\sigma_j}^2}\sum_{i=1}^{n}{\varepsilon_{ij}}^2)$$

Di Marzio, Panzera and Venieri (2015) derive a nonparametric approach to compositional data regression for local constant and local linear estimators.

Let $\mathcal{K}$ be a continuous function with maximum at 0 such that $\mathcal{K}(-u) = \mathcal{K}(u) \geq 0, u \in \mathbb{R}$, and $\int_{\mathbb{R}} \mathcal{K}(x)dx < +\infty$. A D-variate simplicial kernel can be defined, for each $\boldsymbol{u} \in \boldsymbol{S}^D$, as

$$K(\boldsymbol{u}) := \frac{\mathcal{K}(\|\boldsymbol{u}\|_a)}{\int_{S^D}\mathcal{K}(\|\boldsymbol{u}\|_a)d\lambda_a(\boldsymbol{u})}$$

where λ_a stands for the Aitchison measure on $\boldsymbol{S}^D$.

For a general nonparametric model

$$Y_i = m(\boldsymbol{X}_i) + \epsilon_i$$

local constant estimator for $m(\mathbf{x})$ is derived as solution to:

$$\underset{b_0}{\operatorname{argmin}} \sum_{i=1}^{n}\{Y_i - b_0\}^2 K_H(\boldsymbol{X}_i \ominus \boldsymbol{x})$$

This leads to: $\widehat{m}(\boldsymbol{x}; \boldsymbol{H}) = \frac{\sum_{i=1}^{n} K_H(\boldsymbol{X}_i \ominus \boldsymbol{x})Y_i}{\sum_{i=1}^{n} K_H(\boldsymbol{X}_i \ominus \boldsymbol{x})}$

For the local linear estimator, assuming C-differentiability of m at x, $m(\boldsymbol{X}_i)$ can be expanded according to the following first-order Taylor series,

$$m(\boldsymbol{X}_i) \approx m(\boldsymbol{x}) + \mathfrak{D}_m(\boldsymbol{x})\ln(\boldsymbol{X}_i \ominus \boldsymbol{x})$$

and the estimator for $m(\boldsymbol{x})$ is the solution to

$$\underset{\{b_0,b_1\}}{\operatorname{argmin}} \sum_{i=1}^{n}\{Y_i - b_0 - b_1\ln(\boldsymbol{X}_i \ominus \boldsymbol{x})\}^2 K(\boldsymbol{X}_i \ominus \boldsymbol{x})$$

with K the simplicial kernel.

The solution can be expressed as

$$\widehat{m}(x) = i^T(\mathbb{X}^T\mathbb{K}\mathbb{X})^{-1}\mathbb{X}^T\mathbb{K}\mathbb{Y}$$

2 Wavelets for triangles

Our approach to the derivation of wavelets is based on a previous approach of Dey and Wang (2015) – for the presentation in Terrassa we will present the results for the generalization to any simplex space.

Let T be a triangle. Consider its succesive refinement $\{T_{j,k}; j \geq 1, k \in \mathcal{I}^j\}$, where $\mathcal{I}^j = \{1, \dots, 4^j\}$, and each triangle in a finer scale is constructed from one in a coarser level by midpoint subdivision, denoted the

ISBN 978-84-947240-2-2

resulting three subtriangles by

$$T_{j,k} = T_{j+1,k_0} \cup T_{j+1,k_1} \cup T_{j+1,k_2} \cup T_{j+1,k_3}$$

For $d > 0$, let

$$P_d = \{x^i y^j; i + j \leq d\},$$
$$P_d(T) = \{f; f|_T \in P_d, f|_{R^2 \setminus T} = 0\}$$

For $T = T_0 \cup T_1 \cup T_2 \cup T_3$ define $V = P_d(T_0) \oplus P_d(T_1) \oplus P_d(T_2) \oplus P_d(T_3)$ and $W = V \ominus P_d(T)$

We have to construct an orthonormal basis for $L^2(T)$,

$$\{\chi_{T_0}, h_{e^i_{j,k}}; i = 1,2,3, j \geq 0, k \in \mathcal{I}^j\}$$

and any $f \in L^2(T)$ has a decomposition

$$f = \alpha_0 \chi_{T_0} + \sum_{i=1}^{3} \sum_{j=0}^{\infty} \sum_{k \in \mathcal{I}^j} \beta^i_{j,k} h_{e^i_{j,k}}$$

Scaling functions $\chi^l_{T_0}$ are orthogonal polynomials supported by T_0, such that

$$\langle \chi^l_{T_0}, \chi^{l'}_{T_0} \rangle = \delta_{ll'}$$

Constructing the power basis in barycentric coordinates on triangle T_B and applying Gram-Schmidt yields a set of Legendre polynomials $\{\pi^l; l \geq 0\}$ on T_B.

Let $\chi^l_{T_B} = \pi^l 1_{T_B}$. The resulting sequence $S = \{\chi^l_{T_B}\}$ will be a triangular sequence of orthogonal polynomials.

Mutilation gives the scaling functions:

$$\chi^l_{T_{j,k}} = \sqrt{\frac{1}{2|T_{j,k}|}} \chi^l_{T_B}$$

Orthonormal basis for

$$W = \{P_d(T_0) \oplus P_d(T_1) \oplus P_d(T_2) \oplus P_d(T_3)\} \ominus P_d(T)$$

Let $\{\pi^l_T\}$ be the Legendre polynomials mutilated to the triangle T.

Then:

$$h^l_i(\tau_1, \tau_2, \tau_3) = \begin{cases} \pi^l_{T_i}\big((\tau_1, \tau_2, \tau_3) M_{T \to T'}\big) & on\ T_i \\ -\pi^l_{T_i}\big((\tau_1, \tau_2, \tau_3) M_{T \to T'}\big) & on\ T \setminus T_i \\ 0 & otherwise \end{cases}$$

Steps to the orthonormal basis:

1) Orthogonalize $\{h^l_i\}$ against $P_d(T)$ and replace $\{h^l_i\}$ by $h^l_i - \sum_{l=0}^{M-1} \langle h^l_i, \chi^l_T \rangle \chi^l_T$
2) $span(\{h^l_i\}) = P_d(T)^\perp \cap \oplus_{j=0}^{r} P_d(T_i)$
3) Orthogonalizing $span(\{h^l_i\})$ by Gram-Schmidt, we get an orthonormal basis for W

Finally, define the motherwavelets as

$$h_{e^i_{j,k}} = \sqrt{\frac{1}{2|T_{j,k}|}} h^l_i$$

3 Wavelet regression for compositional data

We use the derived wavelet transform to perform the regressions. This allos several possibilities:

- parametric (MNL, Dirichlet and other regular types of CoDa regressions)
- nonparametric (local constant or local linear – Di Marzio et al., 2015)

ISBN 978-84-947240-2-2

- Bayesian (following basic derivations in Shimizu et al., 2015; and van der Merwe, 2018)

Following Dey and Wang (2015) we follow a Bayesian approach using wavelet priors.

We use Multivariate Laplace prior, based on symmetric multivariate Laplace distribution, which has a characteristic function:

$$\varphi(t;\mu,\Sigma) = \frac{\exp(i\mu' t)}{1+\frac{1}{2}t'\Sigma \mathrm{t}}$$

and probability density function: $f_x(x_1,\dots,x_k) = \frac{2}{(2\pi)^{\frac{k}{2}}|\Sigma|^{0.5}}\left(\frac{x'\Sigma^{-1}x}{2}\right)^{\frac{v}{2}} K_v(\sqrt{2x'\Sigma^{-1}x})$, where: $v=(2-k)/2$;

K_v: modified Bessel function of the second kind.

For the case of simplicial-real regression we use the following hyperparameter priors:

$$\beta_{0j} \sim N(a_{0j}, {b_{0j}}^2)$$
$$\boldsymbol{\beta}_j \sim MLap(\mathbf{a}_j, {\mathbf{b}_j}^2)$$
$${\sigma_j}^2 \sim IG(c_j, d_j)$$

Another issue when computing the posterior are intractable marginals (van Gerven et al., 2009). As solution we try a deterministic approximate inference method, namely expectation propagation (EP), see Minka (2004).

The posterior distribution on z given the data y can be written in the factorized form

$$p(z) \propto t_0(z)\prod_i t_i(z)$$

where $t_0(z) \propto N\left(y\middle|Xs,\ \sigma^2 I\right)N(v|0,J^{-1}/\lambda^2)N(u|0,J^{-1}/\lambda^2)$, $t_i(z) = t_i(s_i,u_i,v_i) = N(s_i|0,{u_i}^2+{v_i}^2)$.

Term $t_0(z)$ is a Gaussian function, i.e. can be written as $\exp\left(z^T h_0 - \frac{z^T K_0 z}{2}\right)$.

Using EP we approximate $p(z)$ with $q(z) \propto\ t_0(z)\prod_i \overline{t_i}(z)$ where $\overline{t_i}(z)$ are Gaussian functions as well.

We provide results of some initial simulations, with coverage probabilities and standard errors of parametric and nonparametric methods. Our simulation results are based on 10000 simulated data sets and corresponding 1000 resamples.

Simplicial-real

Data	PAR	LC	LL	WAV-ML	WAV-MG
Log(normal)	0.8703	0.8730	0.8785	0.9196	0.8398
100	0.1765	0.1808	0.1783	0.1808	0.1746
	0.9066	0.8818	0.8964	0.9289	0.8840
200	0.1234	0.1271	0.1281	0.1326	0.1228
	0.9315	0.9204	0.9352	0.9468	0.9150
500	0.0979	0.0876	0.0909	0.1023	0.0977
	0.9438	0.9379	0.9407	0.9538	0.9300
1000	0.0823	0.0850	0.0874	0.0930	0.0814
Dirichlet	0.8687	0.8376	0.8797	0.9029	0.8550
100	0.1708	0.1746	0.1725	0.1748	0.1692
	0.9049	0.8817	0.8977	0.9213	0.9000
200	0.1212	0.1247	0.1254	0.1290	0.1198
	0.9396	0.9200	0.9321	0.9482	0.9240
500	0.0979	0.0962	0.1022	0.0887	0.0922
	0.9465	0.9394	0.9446	0.9551	0.9270
1000	0.0823	0.0859	0.0881	0.0814	0.0838

Simplicial-simplicial

Data	PAR	LC	LL	WAV-ML	WAV-MG
Log(normal)	0.8790	0.9166	0.8609	0.9472	0.8482
100	0.1747	0.1718	0.1801	0.1736	0.1711
	0.9157	0.8906	0.9412	0.9382	0.8840
200	0.1172	0.1335	0.1294	0.1273	0.1203
	0.8849	0.9480	0.9820	0.9941	0.9333
500	0.0989	0.0893	0.0945	0.0982	0.0938
	0.9344	0.8910	0.9877	0.9920	0.9486
1000	0.0782	0.0893	0.0848	0.0902	0.0806
Dirichlet	0.9121	0.8711	0.9061	0.9029	0.8636
100	0.1725	0.1676	0.1639	0.1783	0.1709
	0.9001	0.8905	0.9067	0.9489	0.9270
200	0.1176	0.1185	0.1204	0.1251	0.1234
	0.8926	0.9384	0.9321	0.9198	0.9240
500	0.0999	0.0933	0.1012	0.0843	0.0940
	0.9138	0.9119	0.9635	0.9455	0.9734
1000	0.0807	0.0902	0.0925	0.0838	0.0838

Real-simplicial

Data	PAR	LC	LL	WAV-ML	WAV-MG
Log(normal)	0.9054	0.9075	0.8437	0.8998	0.8397
100	0.1660	0.1683	0.1765	0.1805	0.1677
	0.9431	0.8550	0.9600	0.9570	0.8663
200	0.1196	0.1348	0.1281	0.1311	0.1203
	0.8849	0.9196	0.9816	0.9643	0.8866
500	0.0959	0.0902	0.0955	0.0933	0.0900
	0.8876	0.9266	0.9779	0.9523	0.9391
1000	0.0790	0.0857	0.0882	0.0938	0.0774
Dirichlet	0.9186	0.8363	0.9424	0.9300	0.8376
100	0.1691	0.1592	0.1704	0.1694	0.1675
	0.8982	0.8638	0.8613	0.9679	0.8899
200	0.1211	0.1137	0.1264	0.1289	0.1259
	0.9105	0.9853	0.9694	0.9290	0.9425
500	0.0999	0.0943	0.1032	0.0809	0.0931
	0.9039	0.9074	0.9250	0.8983	0.9928
1000	0.0847	0.0866	0.0916	0.0796	0.0855

4 Conclusion

There are multiple possible ways to extend our work, listed shortly below.

ISBN 978-84-947240-2-2

- Extension to other nonparametric methods (e.g. sieves, splines, extensions of local polynomial and locally adaptive estimators)
- Different types of wavelet constructions
- Different types of wavelet (and other) priors and other types of „second-stage“ modelling
- Generalization to tetrahedrons and arbitrary dimensions
- Semiparametric considerations
- Improved simulation and real application evidence (e.g. geological datasets)

In general, nonparametric and semiparametric models promise an interesting area for research in compositional and complex data analysis in future, both in terms of regression approaches as well as more general derivations of statistical tests.

References

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society. Series B*, 44, pp. 139–177.

Aitchison, J. (2003). *The statistical analysis of compositional data*. Reprinted by The Blackburn Press, New Jersey.

Dey, D.K., Wang, Y. (2015). *Wavelet Modeling of Priors on Triangles*. Working paper, University of Connecticut.

Di Marzio, M., Panzera, A., Venieri, K. (2015). Non-parametric regression for compositional data. *Statistical Modelling*. 15(2), pp. 113–133.

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

Egozcue, J.J., Daunis-i-Estadella, J.,Pawlowsky-Glahn, V., Hron, K., Filzmoser, P. (2011). Simplicial regression. The normal model. *Journal of Applied Probability and Statistics*, 6, pp. 87–108.

Gueorguieva, R., Rosenheck, R., Zelterman, D. (2008). Dirichlet component regression and its applications to psychiatric data. *Computational statistics & data analysis*, 52, pp. 5344– 5355.

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

Martın-Fernandez, J.A., Hron, K., Templ, M., Filzmoser, P., Palarea-Albaladejo, J. (2012). Model-based replacement of rounded zeros in compositional data: Classical and robust approaches. *Computational Statistics & Data Analysis*, 56, pp. 2688–2704.

Martın-Fernandez, J.A., Barcelo-Vidal, C. and Pawlowsky-Glahn, V. (2012). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35, pp. 253–278.

Scealy, J.L., Welsh, A.H. (2011). Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society. Series B*, 73, pp. 351– 375.

Shimizu, T.K.O., Louzada, F., Suzuki, A.K., Ehlers, R.S. (2015). *Modeling Compositional Regression with uncorrelated and correlated errors: a Bayesian approach*. arXiv:1507.00225v1.

Stephens, M.A. (1982). Use of the von Mises distribution to analyse continuous proportions. *Biometrika*, 69, pp. 197–203.

Tsagris, M.T., Preston, S., Wood, A.T.A. (2011). A data-based power transformation for compositional data. In *Proceedings of the 4th Compositional Data Analysis Workshop*, Girona, Spain.

Monitoring robust estimates for compositional data

Valentin Todorov[1]
[1]United Nations Industrial Development Organization (UNIDO), Vienna, Austria; *v.todorov@unido.org*

Summary

In a number of recent articles Riani, Cerioli, Atkinson and others advocate the technique of monitoring robust estimates computed over a range of key parameter values (Cerioli et al., 2018; Riani et al., 2019). Through this approach the diagnostic tools of choice can be tuned in such a way that highly robust estimators which are as efficient as possible are obtained. This approach is applicable to different robust multivariate estimates like S- and MM-estimates, MVE and MCD as well as to the Forward Search in which monitoring is part of the robust method.

Key tool for detection of multivariate outliers and for monitoring of robust estimates are the scaled Mahalanobis distances and statistics related to these distances. However, the results obtained with this tool in case of compositional data might be unrealistic since compositional data contain relative rather than absolute information and need to be transformed to the usual Euclidean geometry before the standard statistical tools can be applied. Several specific transformations have been introduced, but Filzmoser and Hron (2008) show that the transformation with the best properties with respect to robust estimates and keeping invariant the Mahalanobis distances is the *ilr* (isometric log-ratio) transformation.

To illustrate the problem of monitoring compositional data and to demonstrate the usefulness of monitoring in this case we start with a simple example and then analyze a real life data set presenting the technological structure of manufactured exports which, as an indicator of their quality, is an important criterion for understanding the relative position of countries measured by their industrial competitiveness. The analysis is conducted with the R package **fsdaR**, which makes the analytical and graphical tools provided in the MATLAB FSDA library available for R users.

Key words: compositional data, forward search, robust estimates, outliers.

1 Introduction

In many cases the data sets are characterized by multivariate observations (vectors) containing relative contributions of parts to a whole. Examples are geochemical composition of rocks, household budget patterns, time budget, ceramic compositions. A plethora of further examples can be found in Aitchison (1986, 2005) and the hundreds of papers published on this topic. Here I want to point out one of the examples which were the motivation for this contribution. Since 2002 the United Nations Industrial Development Organization (UNIDO) publishes the Competitive Industrial Performance (CIP) Index and accompanying report (Todorov and Pedersen, 2017), see `http://stat.unido.org/cip`. Through this index monitoring the industrial competitiveness of countries will to a great extent reflect how well they manage to adapt to these new challenges and embrace the opportunities. The CIP Index is an essential tool for countries to view and compare their industrial competitiveness with that of others. The CIP Index is composed of eight sub-indicators defined within the framework of three key dimensions that capture different aspects of

ISBN 978-84-947240-2-2

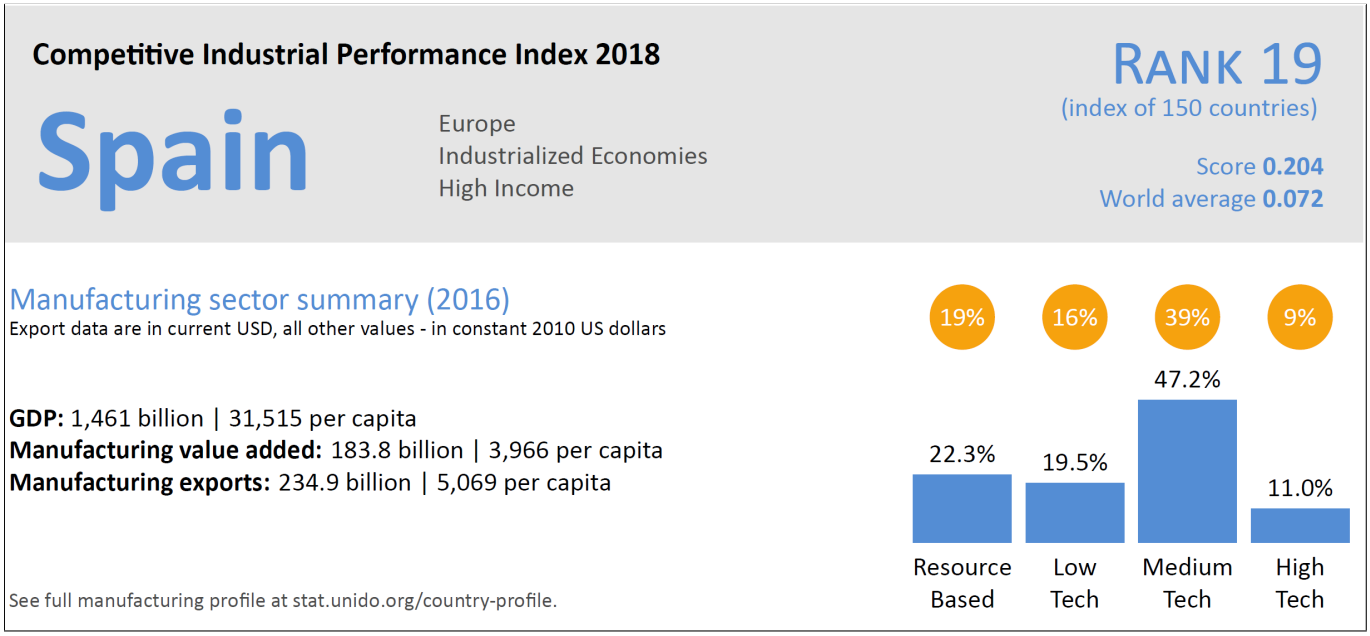


Figure 1: Example of a Competitive Industrial Performance (CIP) Index country profile. The bar chart in the top right corner represents the export structure.

a country's industrial competitive performance. One of these sub-indicators is the technological structure of manufactured exports representing their "quality". There exists an well established decomposition analysis by technology level of the export structure (Lall, 2000) presenting the manufactured exports in four categories: Resource-based, Low technology, Medium technology and High technology (about the source of data and how these categories are defined see Todorov and Pedersen, 2017). Figure 1 is an excerpt from one country profile based on CIP edition 2018. The bar chart in the top right corner represents the export structure of the manufacturing exports of the country. The percentages shown sum up to 100%. However, the country does not export only manufacturing goods, also agricultural, mineral, energy goods or services can compose the total exports. This is shown by the four circles above the corresponding bars — the percentages shown in the circles are the shares of the respective manufacturing category in the total exports. In Section 4.2 is presented the analysis of the export structure as compositional data.

The rest of the paper is structured as follows. Section 2 discusses the need of robust methods and presents briefly the forward search method, the monitoring of different methods and the available software for doing this. In Section 3 the specifics of robust methods in case of compositional data are considered and are illustrated with a simple example. Section 4 presents the monitoring of compositional data on two examples and Section 5 concludes.

2 Forward search and monitoring of robust estimates

In statistical modeling and estimation assumptions like normal distribution or independence are used, however, the practice, is usually different: practical data sets often do not follow these strict assumptions. There might be several different processes inherent in the data generating process, or other effects that cannot be controlled. It is then often unclear how reliable the results are, if the model assumptions are violated. The multivariate aspect of the data used makes the task of outlier identification particularly challenging.

The outliers can be completely hidden in one or two dimensional views of the data. This underlines that univariate outlier detection methods are useless, although they are often favored by researchers because of their simplicity.

Outlier detection and robust estimation are closely related (see Hampel et al., 1986; Hubert et al., 2008) and the following two problems are essentially equivalent:

- Robust estimation: develop statistical techniques which are inherently insensitive to the presence of outliers and find an estimate which is not influenced by these outliers, even if their amount is large (many good robust techniques can tolerate up to 50% contamination). The ability of the estimators to cope with large amount of outliers is measured by their *breakdown point* (*bdp*) which can reach the maximum of 50%. Estimators which can cope with this maximum amount of contamination in the sample are known as *highbreakdown point estimators* (*HBDP estimatos*) and examples of popular HBDP estimatos are Minimum Covariance Determinant (MCD) estimator (Hubert et al., 2017), S-and MM-estimators (Maronna et al., 2006) as well as the Forward Search estimator (Cerioli et al., 2014).
- Outlier detection: find all outliers, which could distort the estimate. A classical approach to detecting multivariate outliers would be to compute the Mahalanobis distance

$$MD(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^t \mathbf{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})} \tag{1}$$

  with $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$ are the sample mean and covariance matrix of the data set $\boldsymbol{X}$ for each $\boldsymbol{x}_i$. Outliers may be identified by large values of $MD(\boldsymbol{x}_i)$. Unfortunately this approach suffers from two problems: (a) *Masking:* multiple outliers can distort the classical estimates of mean $\bar{\boldsymbol{x}}$ and covariance $\boldsymbol{S}$ in such a way (attracting $\bar{\boldsymbol{x}}$ and inflating $\boldsymbol{S}$) that they do not get necessarily large values of $MD(\boldsymbol{x}_i)$ and (b) *Swamping:* multiple outliers can distort the classical estimates of mean $\bar{\boldsymbol{x}}$ and covariance $\boldsymbol{S}$ in such a way that observations which are consistent with the majority of the data get large values of $MD(\boldsymbol{x}_i)$. To cope with these undesirable effects it is necessary to base the diagnostic tools on high breakdown point methods and replace the classical Mahalanobis distances by their robust alternative

$$RD(\boldsymbol{x}_i) = \sqrt{(\boldsymbol{x}_i - \mathbf{T})^t \mathbf{C}^{-1} (\boldsymbol{x}_i - \mathbf{T})} \tag{2}$$

  where $(\mathbf{T}, \mathbf{C})$ is a HBDP robust estimator of multivariate location and scatter.

A solution of the first problem allows us to identify the outliers using their robust residuals or distances while on the other hand, if we know the outliers we could remove or downweight them and then use the classical estimation methods. In many research areas the first approach is the preferred one but often the second one is more appropriate. The focus in the present work is on using robust methods to identify the outliers which can be treated in the traditional way afterwards. In our study we assume that we have a single multivariate population possibly containing outliers. The straightforward approach is to use a 50% estimator, like MCD or S-estimator, however, in case of clean data the results will be with very law efficiency. A first remedy for this is to use reweighting (for MCD) or MM-estimates instead of S-estimates. An alternative approach is the use of adaptive methods based on monitoring a series of fits to the data that indicate good choices of efficiency or bdp (Cerioli et al., 2018; Riani et al., 2019).

The forward search for multivariate analysis is an algorithm for avoiding outliers by recursively constructing subsets of "good" observations, thus providing an automatic form of monitoring. We start by choosing (by some robust criterion) a small subset of observations and then repeatedly extend it in such a way that outliers and other influential observations enter only toward the end of the search, arriving to the final fit that corresponds to the classical statistical estimates. During this process we monitor a suitable diagnostic

measure and the inclusion of outliers is typically signalled by a sharp increase in this measure. Formal description of the forward search method can be found in many papers, see for example Riani et al. (2019). These underlying idea can be extended to many other techniques like S- and MM-estimates. The subsequent estimations are presented in monitoring plots of all n squared Mahalanobis distances which can be combined with brushing to relate Mahalanobis distances to data points exhibited in scatterplot matrices. In this way a straight relationship between statistical results and individual observations is established.

From a practical standpoint in data analysis the availability of such tools and their software implementation is essential ro make their applicability to an wide range of data analysis problems. All the methods discussed in a number of papers on forward search and monitoring are implemented in the *Flexible Statistics and Data Analysis (FSDA)* toolbox (Riani et al., 2012), freely available (for users with a MATLAB license at hand) from `http://rosa.unipr.it`. It features robust and efficient statistical analysis of data sets, not only in multivariate context but also in regression and cluster analysis problems. A downside of the current software implementing monitoring (FSDA) is that it is based on the commercial software MATLAB, which apart from its license costs, is not so appealing to the majority of the statistical community, where R is more widespread. The heart of the monitoring approach is the ability to present the result in a way revealing as much information as possible. While R has advanced graphical capabilities, these graphics are static, do not allow much interactivity and here is the main advantage of using MATLAB for implementing the monitoring functions. Developing the computational algorithms for R would not be a problem and an R package (Atkinson et al., 2006) implementing forward search was available on CRAN in the past. However, the advantages provided by presenting the results visually in interlinked graphs allowing interaction with the user will be missing. Therefore, a possible solution for making the FSDA toolbox available to the R community is not to port the toolbox, but to implement an R interface to a MATLAB engine running in the background. Such a technical solution is made possible by the MATLAB Runtime which allows to run compiled MATLAB applications on computers that do not have MATLAB installed. An R package interfacing to the FSDA toolbox, **fsdaR** is available at CRAN (Todorov and Sordini, 2019). The main challenge in developing this package were the technical issues (creating a Java interface between an R package and a MATLAB toolbox running on the MATLAB Runtime). Additional technical challenge was how to extend a CRAN package with binaries, in this case the compiled Java code, but even more serious challenge is the design and implementation of the interface (the function calls) in a way acceptable for an R user. Formula interface, optional/default arguments to the functions, object orientation, documentation are just several topics presenting differences between MATLAB and R. For example an R user will prefer to call a method `plot()` on an object returned by a function, instead of passing optional arguments `(..., 'plots', 1, ...)` to the function. Similarly, an R user will not be happy to follow strict positioning of the (mandatory) arguments as this is done in MATLAB and will prefer to use the formula interface where appropriate. Colors and color names, line types and other graphical parameters is also an area requiring a lot of effort to make the two languages compatible. Some of these problems still are not solved but an R package implementing the monitoring functionality combined with advanced dynamic graphics is already available at CRAN.

Almost all robust estimation methods are computationally intensive and the computational effort increases with increasing number of observations n and number of variables p towards the limits of feasibility. Since the key idea presented here is to monitor quantities of interest, such as parameter estimates, measures of discrepancy and test statistics, as the model is fitted to data subsets of increasing size, it is inevitable that the computational effort grows exponentially and it is obvious that none of these procedures would be feasible, if special care was not taken in their implementation in FSDA. Riani et al. (2015) describe the efficient routines for fast updating of the model parameter estimates in forward search and show that the new algorithms enable a reduction of the computation time by more than 80% and allow the running time to increase almost linearly with the sample size. In Riani et al. (2014) are given computational advances,

suggesting efficient procedures for calculation of consistency factors of robust S-estimators of location and scale. Todorov (2018) presents a brief study of the computational efficiency of the different monitoring methods and states that it is still an open issue and further work is necessary to make the monitoring of S-estimation for different consecutive values of *bdp* efficient for large data sets.

3 Compositional data and robust methods

Compositional data are multivariate data with strictly positive values that sum up to a constant value (1 or 100 per cent or any other constant), see Aitchison (1986). If not all components of the composition have been analyzed, this constant sum property is not directly visible in the data, but the relation between the components is still constrained. The sample space of compositional data is thus the simplex

$$S^D = \{\mathbf{x} = (x_1, \ldots, x_D)^T, x_i > 0, \sum_{i=1}^{D} x_i = 1\} \tag{3}$$

where the simplex sample space is a $D-1$ dimensional subset of R^D. Standard statistical methods can lead to doubtful results if they are directly applied to original closed data.

Key tool for detection of multivariate outliers and for monitoring in the approach described in the previous Section are the scaled Mahalanobis distances (see Equations 1 and 2) and statistics related to these distances. However, the results obtained with this tool in case of compositional data might be unrealistic. This closeness constraint makes it necessary to first transform the data from the so called simplex sample space to the usual real space. Then standard statistical methods can be applied to the transformed data and the results are back transformed to the original space. One of the most convenient transformation is the family of logratio transformations but it is not clear how the different transformations will affect the Mahalanobis distances used for ranking the data points according to their outlyingness. Filzmoser and Hron (2008) considered three well known transformations and showed how these transformations, namely the additive, the centered and the isometric logratio transformations, will affect the Mahalanobis distances computed by classical and robust methods. They show that in case of classical location and covariance estimators all three transformations lead to the same Mahalnobis distances, however, only *alr* and *ilr* extend this property to any affine equivariant estimator.

To illustrate the problem of applying robust methods to compositional data we start with a simple example based on the data set `Vegetables`. The source of data is `https://ndb.nal.usda.gov/ndb/nutrients/index` and the data set is also available in the R package **easyCODA** (Greenacre, 2019). The data set contains the compositions of protein, carbohydrate and fat as a percentage of their respective totals for ten different vegetables and is shown in Table 1. As pointed out by Filzmoser and Hron (2008) for other compositional data sets, we cannot apply standard outlier detection based on Mahalanobis distances, neither classical nor robust, directly to the data set, because, since it is closed, its covariance matrix is singular. Applying the outlier detection methods from the R package **rrcov** (Todorov and Filzmoser, 2009) as well as the methods from the MATLAB toolbox FSDA (Riani et al., 2012) result in an error. After applying *ilr* transformation the data will be open and the bivariate structure is revealed as shown in the distance-distance plot in Figure 2 (robust Mahalanobis distances computed by MCD are plotted against classical Mahalanobis distances). Four observations, namely mushrooms, carrots, corn and beans, are identified as potential outliers (using the 0.975 quantil of the χ^2 distribution as a cut off). The classical Mahalanobis distance, computed with the sample mean vector and covariance matrix does not identify any outlier.

	Protein	Carbohydrate	Fat
Asparagus	35.66	61.07	3.27
Beans(soya)	42.05	35.88	22.07
Broccoli	48.69	43.78	7.53
Carrots	8.65	89.12	2.23
Corn	11.32	85.70	2.98
Mushrooms	16.78	77.25	5.97
Onions	10.44	88.61	0.95
Peas	26.74	71.29	1.97
Potatoes(boiled)	7.84	91.70	0.46
Spinach	41.57	52.76	5.67

Table 1: `Vegetables data set` from the R package **easyCODA**: composition of three nutrients in ten different vegetables.

Since the (closed) data in this example are three-dimensional they can conveniently be presented in a ternary diagram (right hand panel of Figure 2). To better visualize the multivariate data structure we superimpose 0.975 tolerance ellipses of the Mahalanobis distances computed by the sample mean and covariance (blue) and by MCD (red) respectively. The ellipses are back-transformed to the original space using the inverse *ilr* transformation as proposed in Filzmoser and Hron (2008). The ellipse corresponding to the classical estimates (blue) covers all data points, while the robust one (red), based on MCD excludes the four points identified as potential outliers.

4 Monitoring for compositional data

Now we will illustrate the methods and ideas presented in Sections 2 and 3 on two extensive examples. Both data sets were not analyzed previously in the literature in the context of outlier detection and we do not have any information about the presence of outliers. Therefore we start by the standard outlier detection methods in the R package **rrcov** based on MCD and robust Mahalanobis distances. Of course, these will work only after suitably transforming the data, which we do with the *ilr* transformation. Then we continue with S- and MM-estimation, as well as the Forward search, and the corresponding monitoring functions from the R package **fsdaR**. Finally we demonstrate the brushing and linking functionalities for establishing a straight relationship between statistical results obtained and the individual observations.

4.1 Example 1: `FishMorphology` data set

For our first example of illustrating the problem of monitoring compositional data we use the `FishMorphology` data set from the package **easyCODA** (Greenacre (2019), see also Greenacre and Primicerio (2010)). The data set consists of 26 morphometric measurements, in millimeters, on a sample of 75 fish of the species Arctic charr (*Salvelinus alpinus*. Additionally, to each observation, sex (male of female), habitat (littoral, close to the shore and pelagic, in deeper water far from the shore) and the body mass are recorded. For our example we select only the former habitat (59 observations) and 10 out of the 26 morphometric measures. Further we remove the observation with $ID = 51$ which is an obvious outlier and thus remain with a data set of 58 observations and 10 variables. This data set is not strictly compositional, but can be treated in the same way as compositional if closed by dividing each observation by the corresponding row sum.
Needless to say that neither the classical nor any robust (MCD, S- or MM-) covariance matrix can be used

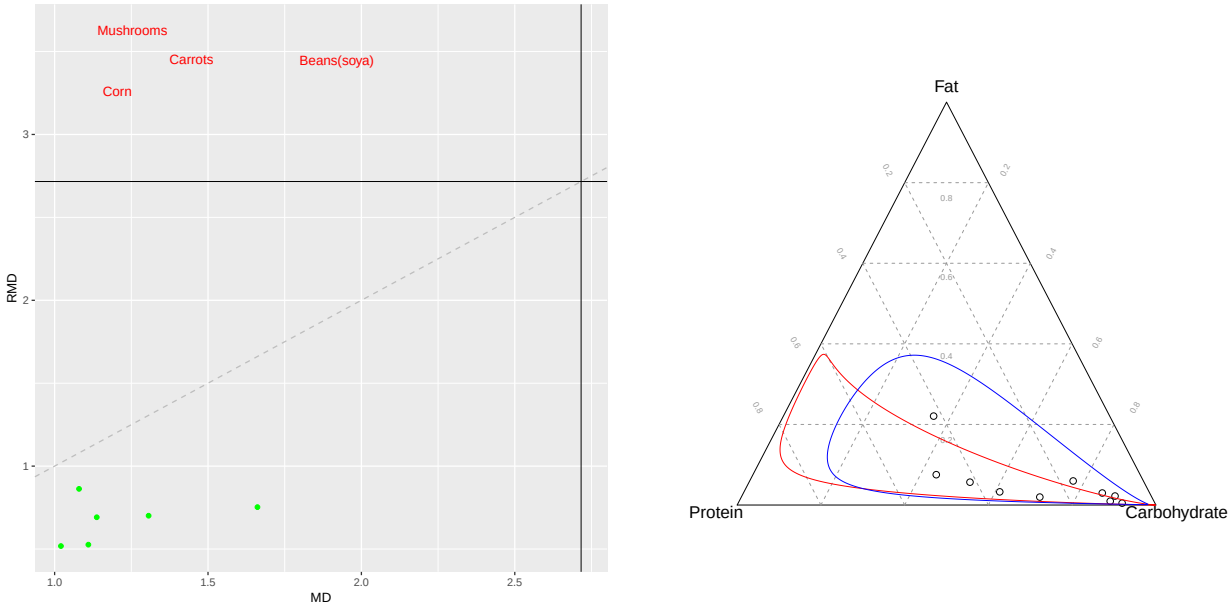


Figure 2: Vegetables data set, *ilr* transformed. MCD distance-distance plot in the left hand panel. A ternary diagram with transformed Mahalanobis distances tolerance ellipses, classical and robust

for computing Mahalanobis distances, since they are singular. After applying *ilr* transformation the data will be open and the multivariate structure is revealed as shown in the distance-distance plot in Figure 3 (robust Mahalanobis distances computed by MCD are plotted against classical Mahalanobis distances). Observations 10, 19 and 54 are identified as potential outliers and observation 47 is a border case (using the 0.975 quantil of the χ^2 distribution as a cut off). The right hand panel of Figure 3 shows the robust and classical chi-square plots which present the squared (robust) Mahalanobis distances against the quantiles of the χ^2_D distribution. Since the (closed) data are in more than three dimensions they cannot be conveniently presented in a ternary diagram as the `Vegetables` data set in the right hand panel of Figure 2. Computing S-estimates with 50% breakdown and Tukey's biweight function (Fig. 4, left panel) produces similar result as the reweighted MCD, identifying only the two outliers but missing the two border cases, however the S-estimates with reduced breakdown point (with the hope to obtain better efficiency) does not identify any outliers (right panel in Fig. 4). Similarly, computing MM-estimates with 80% efficiency and Tukey's biweight function (Fig. 5, left panel) produces similar result as the reweighted MCD, however the MM-estimates with the default efficiency does not identify any outliers (right panel in Fig. 5). As Cerioli et al. (2018) point out, the recommended default efficiency of 95% or 99% for the MM-estimates might be too optimistic, also in our case. Following their approach for data driven balance between robustness and efficiency in the case of compositional data we present in the following the monitoring of the estimation parameters (breakdown point and efficiency) resulting in plots of the squared Mahalanobis distances of the *ilr* transformed data. Figure 6 presents the monitoring of the MM-estimation. The left-hand panel shows the Mahalanobis distances for a series of robust MM fits for subsequent values of the efficiency. These are stable until the efficiency reaches 0.55 when the fit changes abruptly and remains so until 0.85 when it changes again and becomes similar to the maximum likelihood and remains stable until the efficiency reaches 1. It reveals why the index plot of the MM-estimates in Figure 5 did not show any outliers—for efficiency higher than 0.85 the fit is identical to the maximum likelihood. This is also clearly seen from the correlation monitoring in the right hand panel. It shows the monitoring of the three correlation measures which summarize the structure of the plot in the left-hand side by the correlation of the ranks between the squared Mahalanobis distances at adjacent monitoring values. The three standard measures of correlation

 ISBN 978-84-947240-2-2

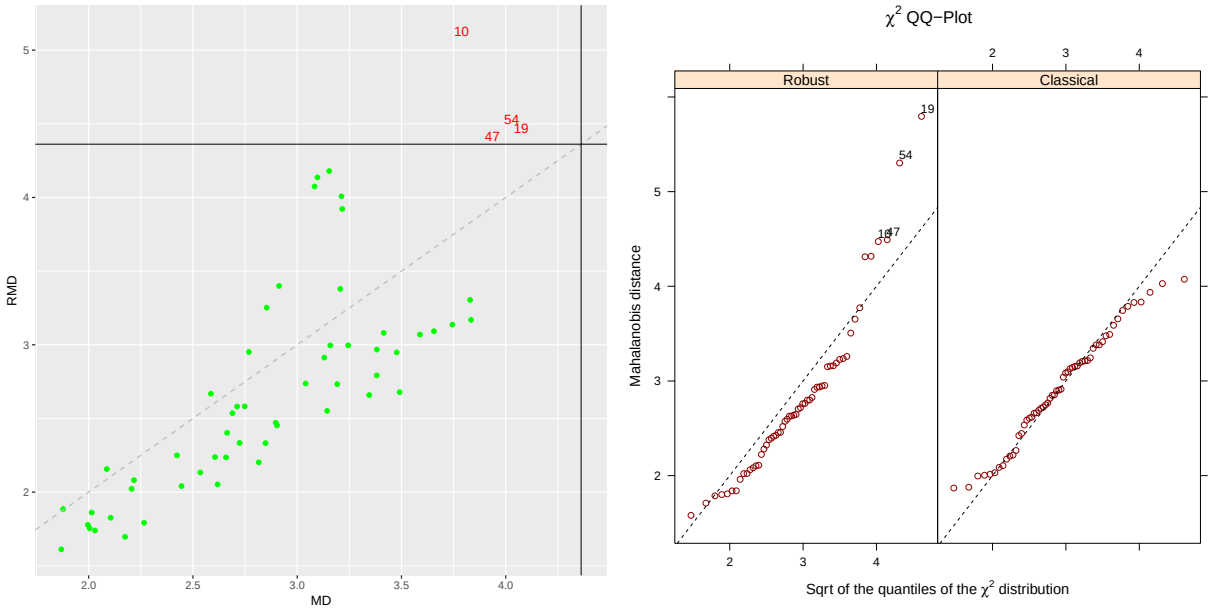


Figure 3: Fish Morphology data set (littoral habitat), *ilr* transformed. MCD distance-distance plot in the left hand panel. A χ^2 QQ plot, classical and robust, in the right hand panel

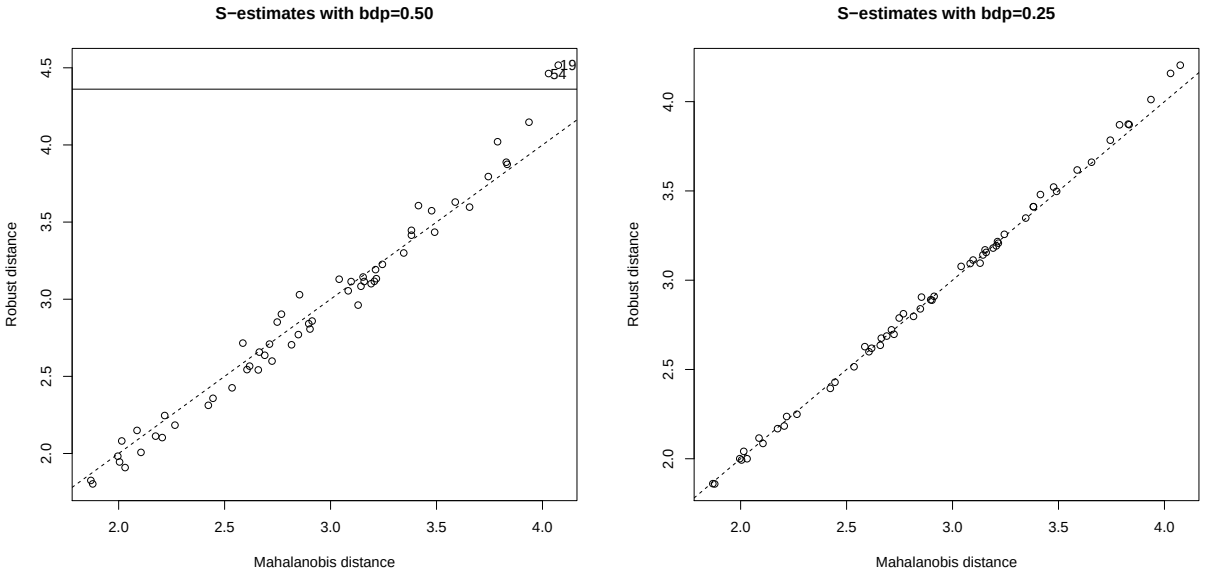


Figure 4: Fish Morphology data set (littoral habitat), *ilr* transformed. S-estimation with breakdown point 50% in the left-hand panel. In the right-hand panel—S-estimation with 25% breakdown.

 ISBN 978-84-947240-2-2

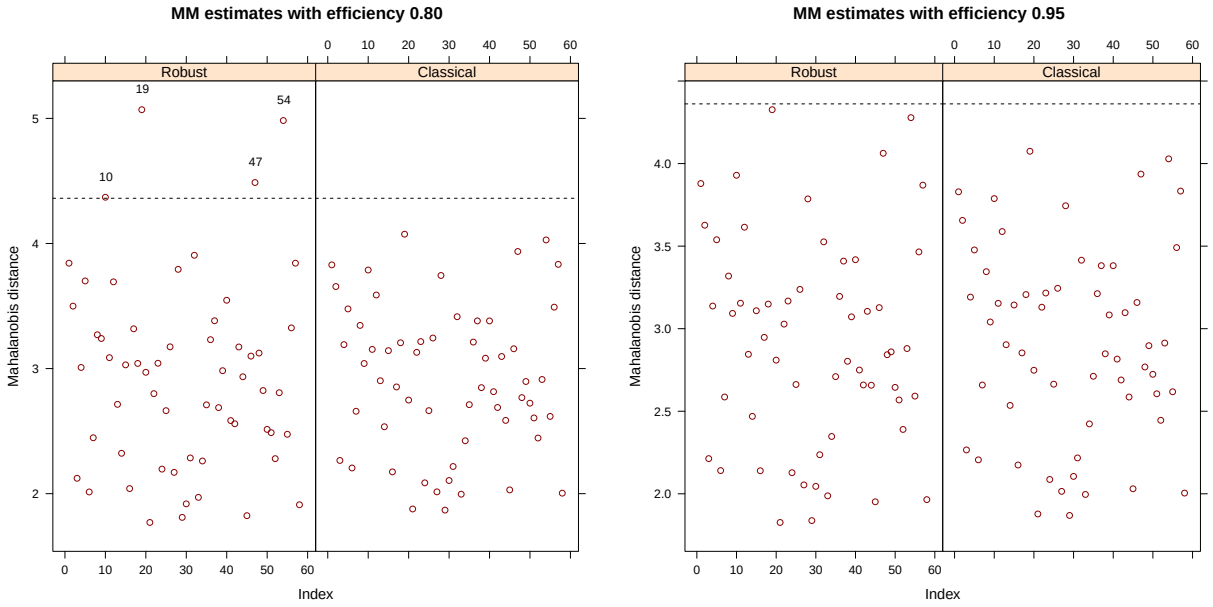


Figure 5: Fish Morphology data set (littoral habitat), *ilr* transformed. MM estimation with efficiency 80% in the left hand panel. In the right hand panel - MM-estimation with 95% efficiency.

are:

1. *Spreaman* This is the correlation between the ranks of the two sets of observations.
2. *Kendal* The concordance of the pairs of ranks.
3. *Pearson* The product-moment correlation coefficient.

All three correlations indicate the abrupt change at 55%. As we have already seen in Figure 5, for `eff=80%` the analysis is still robust but increasing the efficiency to 95% and more results in a non-robust analysis. Using the brushing functionality of the package, we can identify the outlying units, as shown in Figure 7: in the right hand panel the outliers are shown as red circles. A more advanced version of the brushing function is shown in Figure 8. It is possible to do the brushing in several steps, in each selecting different points. The points selected at each step are added to the points selected in the previous step but are presented in different pattern/color. In the first step we select the two outlier which enter the model at around 96% efficiency and they are shown as red circles. In the second step the two outliers that enter the model at around 85% efficiency are selected and they are shown as black stars.

Computing S-estimates with 50% (asymptotic) breakdown point and Tukey's biweight function produces similar result as the reweighted MCD, however, this is not the case if the breakdown point is reduced to say 25%. In Figure 9 is presented the monitoring of the S-estimation. As we have already seen in Figure 4, for `bdp=50%` the analysis is robust but reducing the breakdown to say 25% (with the hope to increase the efficiency) results in a non-robust analysis. This is clearly seen in the left hand panel of Figure 9 although not that clear in the correlation plot on the right side.

4.2 Example 2: Technology intensity of exports

The technological structure of manufactured exports as an indicator of their "quality" is an important criteria for understanding the relative position of countries measured by their industrial competitiveness and

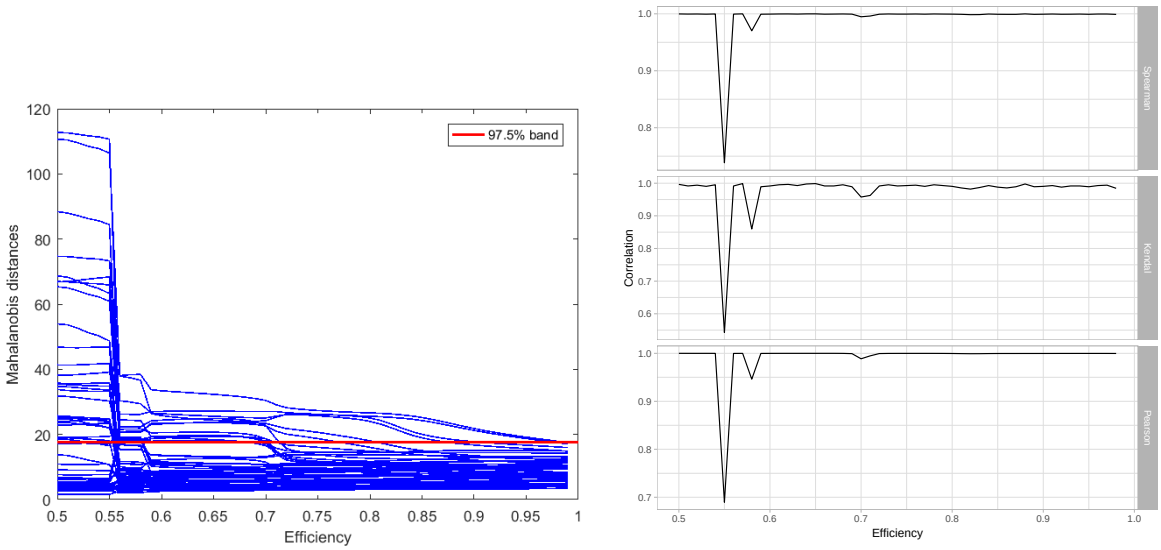


Figure 6: Fish Morphology data set (littoral habitat), *ilr* transformed. The left-hand panel shows the squared Mahalanobis distances from monitoring MM-estimation and the right-hand panel—the correlation between distances for consecutive values of eff (efficiency)

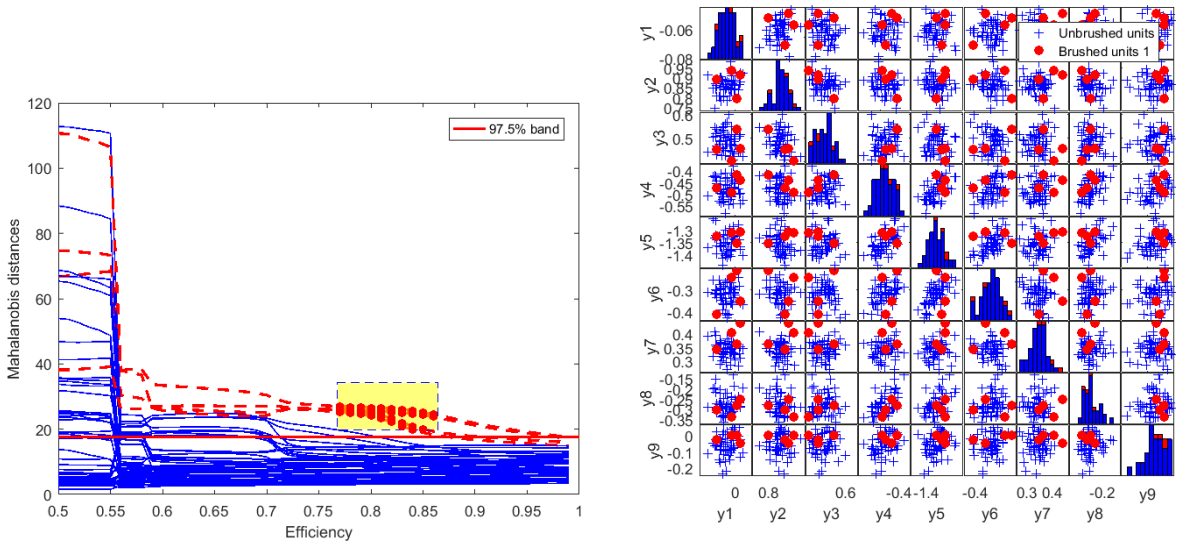


Figure 7: Fish Morphology data set (littoral habitat), *ilr* transformed. The left-hand panel shows brushing of the monitoring plot of MM-estimation and the right-hand panel—the scatter plot matrix of the units, identifying the four outliers.

 ISBN 978-84-947240-2-2

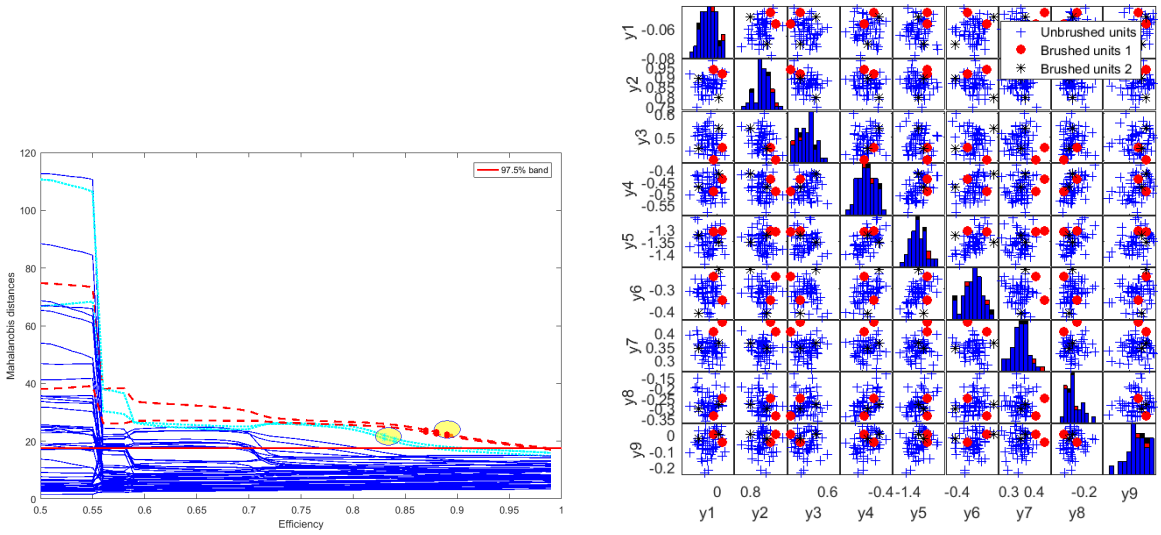


Figure 8: Fish Morphology data set (littoral habitat), *ilr* transformed. Brushing in several steps: the left-hand panel shows brushing of the monitoring plot of MM-estimation and the right-hand panel—the scatter plot matrix of the units, identifying the four outliers. The two which enter the model at around 96% are shown as red circles and those that enter at around 85% - as black stars

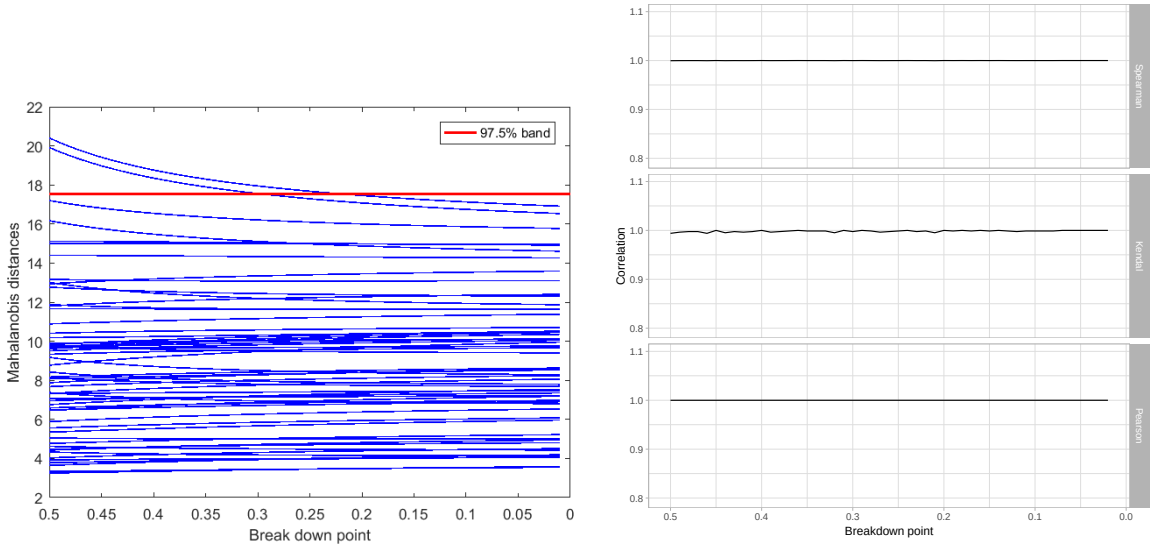


Figure 9: Fish Morphology data set (littoral habitat), *ilr* transformed. The left-hand panel shows the squared Mahalanobis distances from monitoring S-estimation and the right-hand panel—the correlation between distances for consecutive values of *bdp* (breakdown)

 ISBN 978-84-947240-2-2

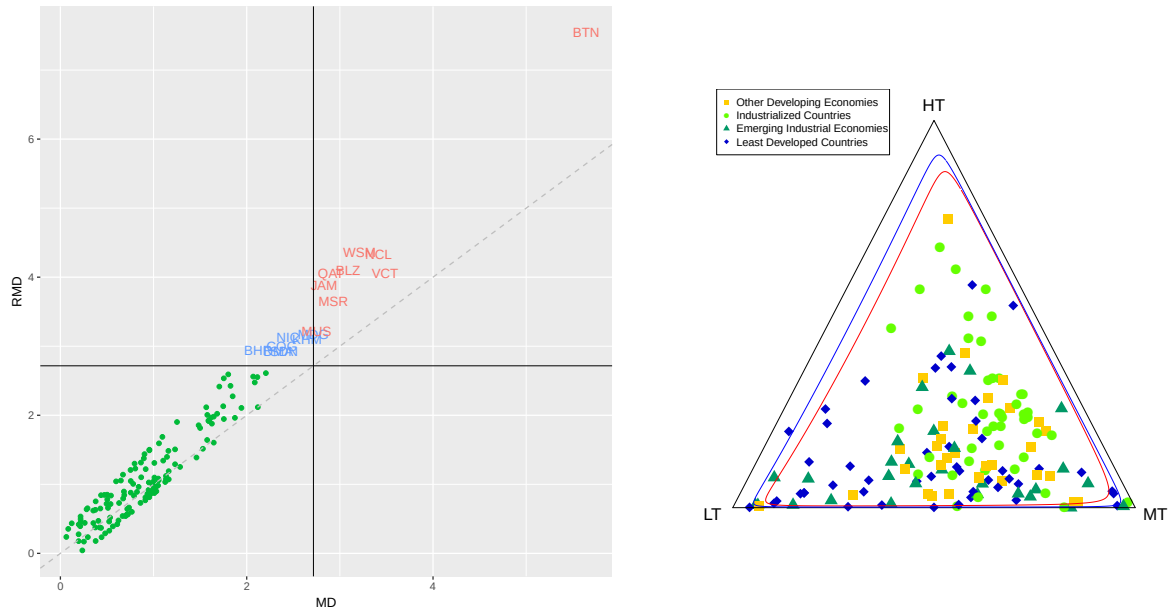


Figure 10: Technological structure of manufactured exports, *ilr* transformed. MCD distance-distance plot in the left hand panel. A ternary diagram with transformed Mahalanobis distances tolerance ellipses, classical and robust.

the determinants of the competitive ability, which are particularly reflected in changes to manufacturing value added and manufactured exports (Todorov and Pedersen, 2017). There exists an well established decomposition analysis by technology level of the export structure (Lall, 2000) presenting the manufactured exports in four categories: Resource-based, Low technology, Medium technology and High technology (about the source of data and how these categories are defined see Todorov and Pedersen, 2017). The data set is available in the R package **rrcov3way** (Todorov, 2017).

For our example we select only one year, 2012 and remove any countries with missing data, remaining with 153 observations. Needles to say that applying the outlier detection methods from the R package **rrcov** or the methods from the MATLAB toolbox FSDA to the original data are meaningless: the reweighted MCD, for example, identifies 79 outliers out of 153 observations. After applying *ilr* transformation the data will be open and the structure is revealed as shown in the distance-distance plot in Fig 10. Now 22 observations are identified as outliers by the reweighted MCD estimator.

This is definitely a compositional data set (the four categories are parts of one whole) but the closure is not visible when inspecting the row sums. This is due to the fact that we consider only the manufactured exports while the countries also export agricultural, mining and other products. This demonstrates the problem of the so called *subcompositions* (Aitchison, 1986)—we cannot hope that the effect of the closure will disappear if not all parts are included in the analysis and an appropriate transformation is needed.

We continue by running the automatic outlier detection procedure based on forward search. As visible in the left hand panel of Fig. 11 the signal is at observation 107, indicating that it and the succeeding observations might be outliers. Resuperimposition of envelopes leads to the identification of 29 outliers (which turn out to be identical to the outliers detected by the raw (not reweighted) MCD). Performing the same analysis on the original data (not shown here) indicates a signal at observation 93 and and identifies 61 observations as outliers.

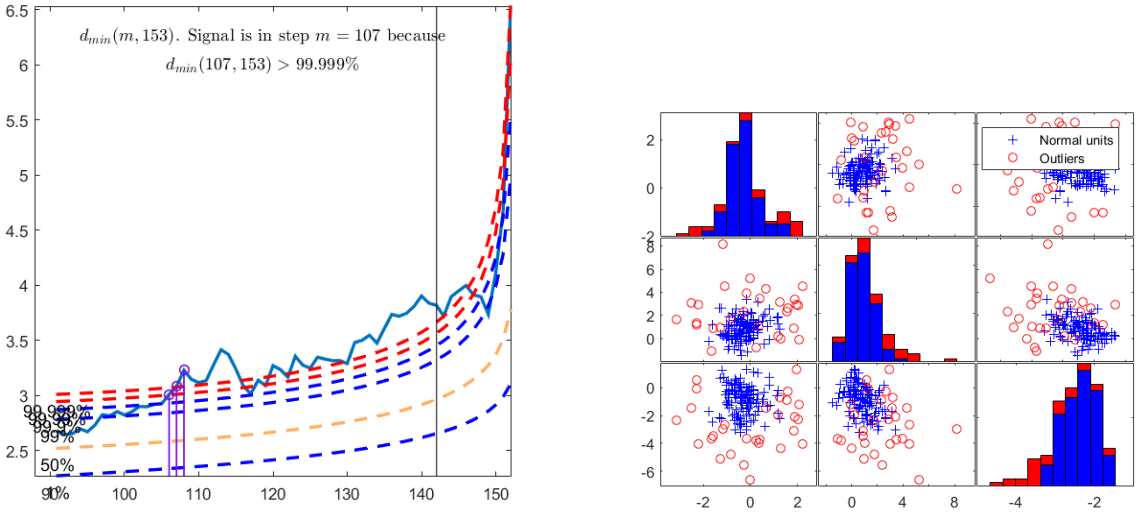


Figure 11: Technological structure of manufactured exports, *ilr* transformed. The left-hand panel shows the forward search plot of minimum Mahalanobis distance, with a signal for the presence of outliers. The right hand panel shows the scatter plot of the data with the 29 observations identified as outliers by FS as red circles.

Fig. 12 shows the monitoring of the Mahalanobis distances of the S- and MM-estimation. The S-estimator with 0.5 bdp is similar to the maximum likelihood. Not much difference is shown in the monitoring plot of the MM-estimation.

5 Conclusions

Robust methods are not only useful but also a required tool when analyzing real life data which very often are plagued by the presence of outliers. It does not matter if the robust methods are used directly to fit models or indirectly to identify outliers, some arbitrarily chosen parameters can have a destructive effect on the results. In a number of recent articles Riani, Cerioli, Atkinson and others advocate the technique of monitoring robust estimates computed over a range of key parameter values. Through this approach the diagnostic tools of choice can be tuned in such a way that highly robust estimators which are as efficient as possible are obtained. This approach is applicable to different robust multivariate estimates like S- and MM-estimates, MVE and MCD as well as to the Forward Search in which monitoring is part of the robust method. We show that in order to apply these adaptive methods to compositional data which are parts of some whole and in most cases they are recorded as closed data, i.e. data summing to a constant, such as 100%, it is necessary to transform the compositional data. Using a suitable transformation, the key measure for detecting outliers, the Mahalanobis distance, remains invariant and allows to successfully apply all methods that were initially developed for open data. We demonstrate on several examples how the monitoring can be conducted, providing highly efficient estimates and demonstrate the role of advanced dynamic graphics like brushing and linking for establishing a straight relationship between statistical results and individual observations. All computations were performed with the **fsdaR** package available at CRAN which brings almost all the functions of the MATLAB toolbox FSDA to the R user. Since the scope of these functions cover also robust regression (Riani et al., 2014) and robust clustering (Riani et al., 2019) a natural

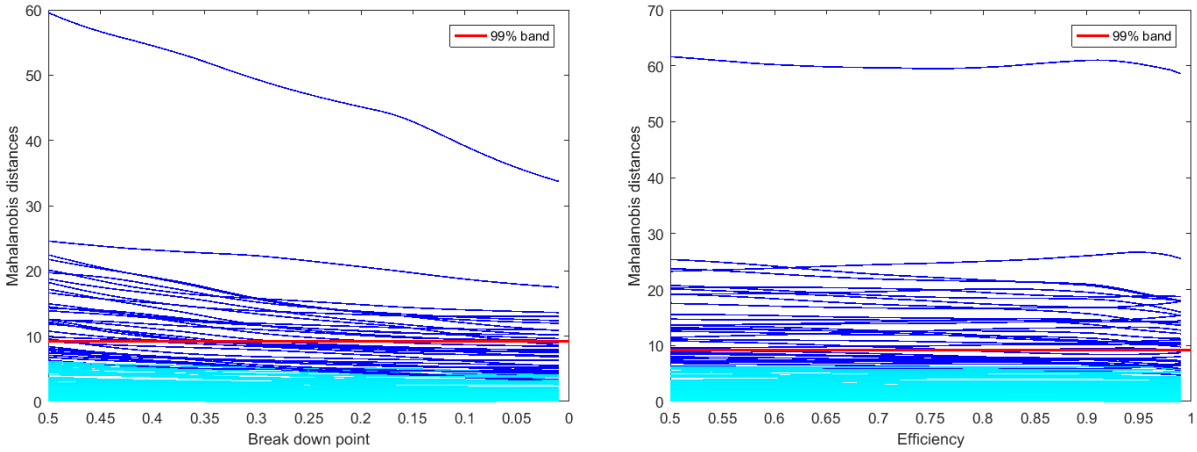


Figure 12: Technological structure of manufactured exports, *ilr* transformed. The left-hand panel shows the squared Mahalanobis distances from monitoring S-estimation and the right-hand panel—from monitoring the MM-estimation.

extension of this study is to consider monitoring for robust regression and clustering for compositional data in the future.

Acknowledgements

The views expressed herein are those of the authors and do not necessarily reflect the views of the United Nations Industrial Development Organization (UNIDO).

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.

Aitchison, J. (2005). *A Concise Guide to Compositional Data Analysis*. Lecture Notes. Available online at CoDaWeb.

Atkinson, A., A. Cerioli, and M. Riani (2006). *Rfwdmv: Forward Search for Multivariate Data*. R package version 0.72-2.

Cerioli, A., A. Farcomeni, and M. Riani (2014). Strong consistency and robustness of the forward search estimator of multivariate location and scatter. *Journal of Multivariate Analysis 126*, 167—-183.

Cerioli, A., M. Riani, A. Atkinson, and A. Corbellini (2018). The power of monitoring: How to make the most of a contaminated multivariate sample (with discussion). *Statistical Methods and Applications 27*, 559–587.

Filzmoser, P. and K. Hron (2008). Outlier detection for compositional data using robust methods. *Mathematical Geosciences 40*, 233—-248.

Greenacre, M. (2019). *Compositional Data Analysis in Practice*. Chapman & Hall / CRC Press.

Greenacre, M. and R. Primicerio (2010). *Multivariate Analysis of Ecological Data*. BBVA Foundation, Bilbao.

Hampel, F., E. Ronchetti, P. Rousseeuw, and W. Stahel (1986). *Robust Statistics. The Approach Based on Infuence Functions*. John Wiley & Sons.

Hubert, M., M. Debruyne, and P. J. Rousseeuw (2017). Minimum covariance determinant and extensions. *WIREs computational statistics*.

Hubert, M., P. J. Rousseeuw, and S. van Aelst (2008). High-breakdown robust multivariate methods. *Statistical Science 23*, 92–119.

Lall, S. (2000). The technological structure and performance of developing country manufactured exports, 1985-98. *Oxford development studies 28*(3), 337–369.

Maronna, R. A., D. Martin, and V. Yohai (2006). *Robust Statistics: Theory and Methods*. New York: John Wiley & Sons.

Riani, M., A. Atkinson, A. Cerioli, and A. Corbellini (2019). Efficient robust methods via monitoring for clustering and multivariate data analysis. *Pattern Recognition 88*, 246–260.

Riani, M., A. Cerioli, A. Atkinson, and D. Perrotta (2014). Monitoring robust regression. *Electronic Journal of Statistics 8*, 646–677.

Riani, M., A. Cerioli, and F. Torti (2014). On consistency factors and efficiency of robust S-estimators. *TEST 23*, 356—-387.

Riani, M., D. Perrotta, and A. Cerioli (2015). The forward search for very large datasets. *Journal of Statistical Software, Code Snippets 67*(1), 1–20.

Riani, M., D. Perrotta, and F. Torti (2012). FSDA: A MATLAB toolbox for robust analysis and interactive data exploration. *Chemometrics and Intelligent Laboratory Systems 116*, 17–32.

Todorov, V. (2017). **rrcov3way***: Robust Methods for Multiway Data Analysis, Applicable also for Compositional Data*. R package version 0.1-10.

Todorov, V. (2018). Discussion of "The power of monitoring: How to make the most of a contaminated multivariate sample" by Andrea Cerioli, Marco Riani, Anthony C. Atkinson and Aldo Corbellini. *Statistical Methods & Applications 27*(4), 631–639.

Todorov, V. and P. Filzmoser (2009). An object oriented framework for robust multivariate analysis. *Journal of Statistical Software 32*(3), 1–47.

Todorov, V. and A. L. Pedersen (2017). Competitive industrial performance report 2016. Volumes I and II. Report, United Nations Industrial Development Organization (UNIDO), Vienna.

Todorov, V. and E. Sordini (2019). **fsdaR***: Robust data analysis through monitoring and dynamic visualization*. R package version 0.4-6.

On machine learning algorithms and compositional data

R. Tolosana-Delgado[1], H. Talebi[2] ,M. Khodadadzadeh[1], and K.G. van den Boogaart[1]
[1]Helmholtz Zentrum Dresden-Rossendorf,
Helmholtz Institute Freiberg for Resource Technology,
Freiberg, Germany; r.tolosana@hzdr.de
[2]CSIRO, Perth Australia

Abstract

Predictive methods such as Lasso regression, partition trees and random forests (RF), artificial neural networks (ANN) and deep learning, or support-vector machines (SVM) and other kernel methods have become in the last years increasingly popular, also in the compositional data community. However, most of the contributions using machine learning algorithms on compositional data just applied the relevant method to an additive, centered or isometric log-ratio (alr, clr, ilr) transformed version of the training data, without caring about the properties of the construct. In this contribution we briefly review the fundamental construction of these methods, and check in which way can they be tweaked or adapted to account for the compositional scale of the data.

As an example, a binary partition tree aims at constructing a hierarchy of classification, where each branch splits the data in two subgroups according to the one single covariable that provides highest purity of the two resulting subgroups; at the end of the hierarchy, all branches contain only data from one pure group. Random Forests Breiman (2001) were introduced to deal with the obvious over-fitting of partition trees, with a double randomisation strategy: first bootstrapping the number of observations, creating B different trees that form the forest; second, each branching of each tree is based not on the whole set of variables, but on a different random subset of them. The fact that at each branching only one variable is actively used makes the method non-invariant under the choice of possible log-ratio transformations. A way to allow for this one feature selection while keeping the relative nature of compositional information would be to build the trees on the set of pairwise log-ratios (pwlr). This applies to all kinds of tree-based methods with compositional covariables.

Key words: affine equivariance, subcompositional coherence, variable selection.

1 Introduction

Predictive methods such partition trees and random forests (RF), artificial neural networks (ANN) and deep learning, or support-vector machines (SVM) and other kernel methods have become in the last years increasingly popular, also in the compositional data community. However, most of the contributions using machine learning algorithms on compositional data just applied the relevant method to an additive, centered or isometric log-ratio (alr, clr, ilr) transformed version of the training data, without caring about the properties of the construct. In this contribution we briefly review the fundamental construction of these methods, and check in which way can they be tweaked or adapted to account for the compositional scale of the data.

After summarizing the most relevant ways of representing compositional data, the paper devotes a section to each family or group of machine learning algorithms. Each of these sections very briefly report what the method does, and several considerations on how to adapt them to compositional data.

2 Compositional data: ratios, logratios and isometric representations

A compositional data set is said to contain only relative information, which can be captured in the form of ratios or log-ratios (Aitchison, 1986). Aitchison (1997) highlighted as well the importance of subcompositions, as the counterpart of marginals for compositional data. For a D-part composition $\mathbf{x} = [x_1, x_2, \ldots, x_D]$, the subcomposition of the set of K parts $\{s_1, s_2, \ldots s_K\} \equiv S$ will be denoted as $\mathbf{x}_S = [x_{s_1}, x_{s_2}, \ldots, x_{s_K}]$.

Several ratios and log-ratios have been used in the literature to extract that relative information. Perhaps the first and simplest one was the closure, i.e.

$$\mathcal{C}[\mathbf{x}] = \left[\frac{x_1}{t(\mathbf{x})}, \frac{x_2}{t(\mathbf{x})}, \cdots, \frac{x_D}{t(\mathbf{x})}\right], \qquad \text{with} \qquad t(\mathbf{x}) = x_1 + x_2 + \cdots + x_D. \tag{1}$$

Before the work of Aitchison, another common way of treating compositions was through simple ratios (or logratios), e.g. via expressions such as

$$x^*_{ij} = \frac{x_i}{x_j}, \qquad \text{or} \qquad \xi_{ij} = \ln x^*_{ij}. \tag{2}$$

The pairwise logratio transformation $\text{pwlr}(\mathbf{x}) = [\xi_{ij}; i < j = 1, 2, \ldots, D]$ was defined by Aitchison (1986), as well as the centered logratio transformation

$$\text{clr}(\mathbf{x}) = \ln\left[\frac{x_1}{g(\mathbf{x})}, \frac{x_2}{g(\mathbf{x})}, \cdots, \frac{x_D}{g(\mathbf{x})}\right], \qquad \text{with} \qquad g(\mathbf{x}) = \sqrt[D]{x_1 x_2 \cdots x_D}, \tag{3}$$

with the logarithm applied component-wise. Another transformation found in this work is the additive logratio transformation,

$$\text{alr}(\mathbf{x}) = \ln\left[\frac{x_1}{x_D}, \frac{x_2}{x_D}, \cdots, \frac{x_{D-1}}{x_D}\right]. \tag{4}$$

Aitchison (1986) introduced as well a notion of compositional distance between two vectors $\mathbf{x}$ and $\mathbf{y}$, as

$$d^2_A(\mathbf{x}, \mathbf{y}) = \frac{1}{D^2}\sum_{i<j}^{D}\left(\ln\frac{x_i}{x_j} - \ln\frac{y_i}{y_j}\right)^2,$$

also based on ratios. Later, with the establishment of the Aitchison geometry (Billheimer et. al. , 2001; Pawlowsky-Glahn and Egozcue, 2001), it was realized that the clr transformation represents an isometry, $d^2_A(\mathbf{x}, \mathbf{y}) \equiv d^2(\text{clr}(\mathbf{x}), \text{clr}(\mathbf{y}))$, being $d(\cdot, \cdot)$ the conventional sum-of-squares, Euclidean distance. Finally, Egozcue et al. (200X) introduced the isometric logratio transformation (ilr) as a Gramm-Schmidt orthogonalisation of the clr coefficients,

$$\text{ilr}(\mathbf{x}) = \text{clr}(\mathbf{x}) \cdot \mathbf{V}, \quad \text{with} \quad \mathbf{V} \cdot \mathbf{V}^t = \mathbf{1}_D - \frac{1}{D}\mathbf{1}_{D\times D} \quad \text{and} \quad \mathbf{V}^t \cdot \mathbf{V} = \mathbf{I}_{D-1}, \tag{5}$$

being $\mathbf{1}_D$ and $\mathbf{1}_{D\times D}$ resp. the $(D \times D)$-identity matrix and $(D \times D)$-ones matrix.

Closed compositional data pose the problem that the scores obtained for each variable depend on all other variables available at that observation. So, if different subcompositions are available, closed values would be not comparable anymore. The same happens with centered logratios. On the other hand, simple ratios, pairwise logratios, additive logratios and isometric logratios do not depend on which subcomposition was used, and they simply produce missing values in some of their coefficients if one or more components are lost. This is a very important issue in machine learning, because the training data fed to the algorithms is not necessarily complete, and ensuring consistency becomes a need.

ISBN 978-84-947240-2-2

3 Partition trees and random forests

Decision trees involve hierarchical segmentation of the predictor space into several simple regions. At each segmentation one single predictor that provides the highest purity of the two resulting regions is selected. To make a prediction for a given observation, we use the mode of the training observations in the region to which that observation belongs. Ensemble tree-based learners, such as bagging, boosting, and random forests, were introduced to deal with the high variance (overfitting) of decision trees and to improve prediction accuracy by generating multiple trees which are then combined to yield a single consensus prediction (Hastie et al, 2009). In the case of compositional predictors, the fact that at each segmentation only one predictor is actively used makes the tree-based methods non-invariant under the choice of possible log-ratio transformations. Indeed, using different log-ratio transformations leads to different predictor spaces and consequently different tree-based predictive models (potentially with different prediction accuracy).

To illustrate the effects of log-ratio transformations on tree-based learners (Random Forests, Breiman (2001), in this implementation) the multi-element near-surface geochemical compositions (compositional predictors) from the National Geochemical Survey of Australia (de Caritat and Cooper, 2011) are used to predict the exposed to deeply buried major crustal blocks (categorical response) of the Australian continent (Talebi et al, 2018). Samples of different sizes (10, 20, 30, 40, and 50) were taken randomly (without replacement) from geochemical components (hundred samples for each size). Each sample was closed (Eq. 1) or transformed via different log-ratio transformations (Eqs. 2-5). For each sample five random Forests classifiers were trained (using raw components, clrs, ilrs, pwlrs, and a combination of raw components plus all the log-ratios as input predictors). Figure 1 shows the distribution of Out-of-Bag (OOB) error estimation for each sample size and log-ratio transformation. As the number of component increases classifiers show more accuracy; however, pairwise log-ratios outperformed other options. The superiority of pwlr is clearer when more components are used to build the classifiers. Combining pwlrs with the other log-ratios does not improve the performance of the classifier and makes the interpretation of the predictor space more complicated. High-dimensionality of pwlrs is well addressed by tree-based predictive methods since they are working with subsets of predictors. In the case of compositional predictors, pairwise log-ratio transformation is recommended as a first choice to train a tree-based predictive model due to their ease of interpretation and superior performance. A recursive feature elimination with resampling technique may further improve the accuracy of the tree-based predicative models trained from pwlrs (Talebi et al, 2018).

4 Some remarks

This superior performance of pwlr on random forests (and in general, partition trees) does not apply to all machine learning methods. Regression methods (linear regression, logistic regression) can be proven to be affine equivariant, namely to produce the same predictions for every logratio transformation (alr, ilr, and even clr or pwlr if the appropriate inversion is used). Ridge regression penalizes the regression goodness of fit (least squares or deviance) by means of the square norm of the regression coefficients, in which case ilr appears to be necessary. Support vector machines, establishing a classifier based on distances, may be required to be applied on isometric representations of the data (ilr, clr but also pwlr). Finally, neural networks require further research, but preliminary results suggest that some implementations of the estimation algorithms may not be affine equivariant, in which case practitioners should use them carefully. These considerations apply to methods using compositional data on the role of predictors.

References

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

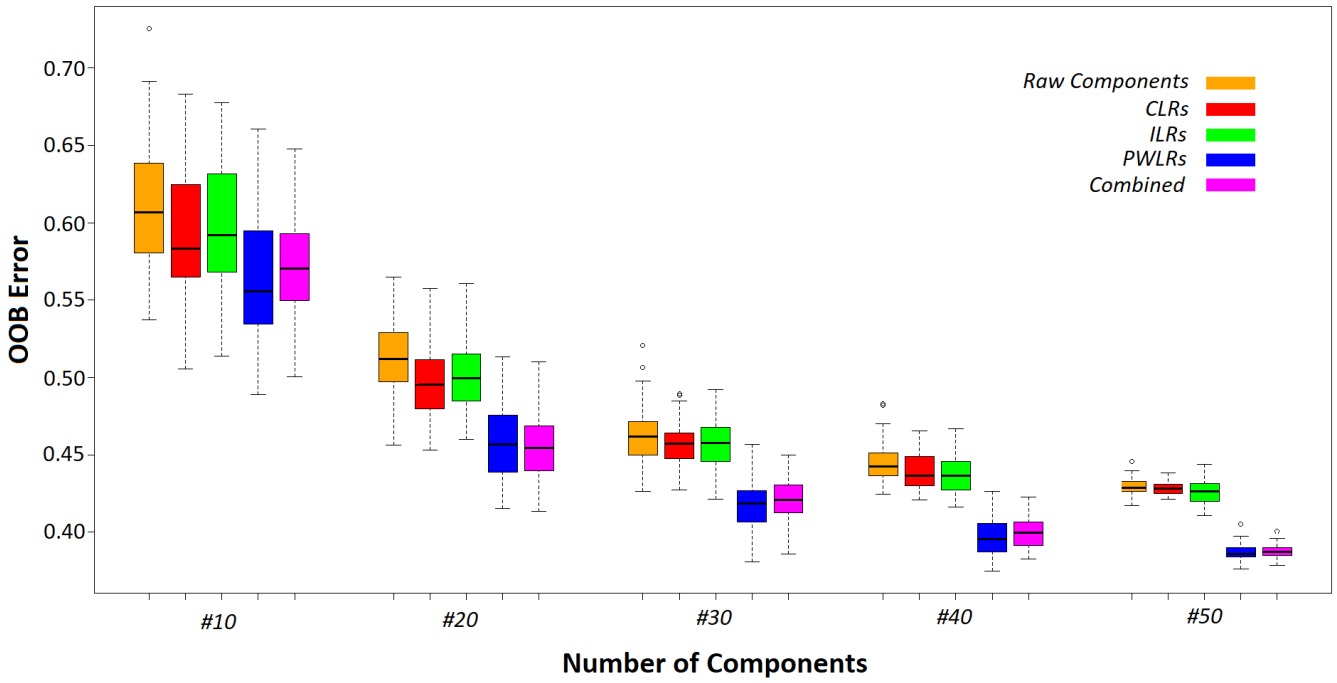


Figure 1: Out-of-bag error estimation for different logratio representations and sample sizes

Press)

Aitchison J (1997) The one-hour course in compositional data analysis or compositional data analysis is simple. In: Pawlowsky-Glahn V (Ed.) *Proceedings of IAMG'97 – The III annual conference of the International Association for Mathematical Geology*, pp 3-35

Billheimer, D. and Guttorp, P. and Fagan, W.F (2001) Statistical interpretation of species composition. *Journal of the American Statistical Association* 456(96):1205–1214

Breiman, L (2001) Random Forests. *Machine Learning* 45:5–32

Caritat P de, Cooper M (2011) *National geochemical survey of Australia: The geochemical atlas of Australia* Geoscience Australia, Record 2011/20

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

Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning.* Springer, New York.

Talebi H, Mueller U, Tolosana-Delgado R, et al (2018). Surficial and Deep Earth Material Prediction from Geochemical Compositions. *Natural Resources Research* (online first) doi: 10.1007/s11053-018-9423-2

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

Pawlowsky-Glahn V, Egozcue JJ (2002). BLU estimators and compositional data. *Mathematical Geology* 34(3):259–274

ISBN 978-84-947240-2-2

Statistical models for point-counting data

P. Vermeesch[1]

[1]Department of Earth Sciences, University College London, UK. *p.vermeesch@ucl.ac.uk*

Abstract

The mineralogical composition of sediment can be estimated by tallying the relative abundances of randomly selected grains. Similarly, the fossil content of a deep sea sediment core may be characterised by tabulating the relative abundances of various species among randomly selected specimens. Or palaeobiological environments may be reconstructed by tabulating the relative frequency of different types of pollen in a palaeosol or charcoal. These are all examples of point-counting experiments. Although the objective of such experiments is to constrain compositions, point-counting data do not exactly fit the mould of traditional compositional data analysis methods. This contribution makes the case that point-counts represent a separate data class that combines elements of compositional data with multinomial statistics.

There exist two key differences between point-counts and compositional data *sensu stricto*. First, point-counting data are sample size dependent, whereas compositional data are not. Sample size matters for point-counting data because the precision of such data increases with sample size. This sample size dependence is not captured by conventional compositional data analysis methods. Second, as a consequence of the sample size dependence, point-counting data frequently include zero values. These are incompatible with the logratio transformations that are used to 'free' compositional data from the constraints of the simplex. This problem may be circumvented by replacing the zeros with small non-zero values (Martín-Fernández et al., 2003). But such 'imputation' methods are a workaround rather than a real solution, and do not cope well with datasets that contain lots of zeros.

Vermeesch (2018) proposed a different solution to the point-counting conundrum that is based on Galbraith (2005)'s solution to a very similar problem in fission track geochronology. The simplest case of a point-counting experiment involves the relative abundances of two components a and b, say. Suppose that the true proportions of these two components follow a logistic normal distribution. If $f[a,i]$ is the fraction of component a in sample i and $1-f[a,i]=f[b,i]$ is the fraction of component b in sample i $(1 \leq i \leq m)$, then

$$\beta[i] = \ln\left(\frac{f[a,i]}{1-f[a,i]}\right) \tag{1}$$

is drawn from a normal distribution with mean μ and standard deviation σ. For a given value of $f[a,i]$, the number of counts of component a and b ($n[a,i]$ and $n[b,i]$, respectively) follow a binomial distribution:

$$p\left(n[a,i], n[b,i] | f[a,i]\right) = \binom{n[a,i]+n[b,i]}{n[a,i]} f[a,i]^{n[a,i]} (1-f[a,i])^{n[b,i]} \tag{2}$$

Combining Equations 1 and 2, we obtain a random effects model for μ and σ, whose likelihood function is given by:

$$\mathcal{L} = \sum_{i=1}^{m} \ln\left\{ \binom{n[a,i]+n[b,i]}{n[a,i]} \int_{-\infty}^{\infty} \frac{\exp\left(\beta n[a,i]\right)}{\left[\exp\left(\beta\right)+1\right]^{n[a,i]+n[b,i]}} \frac{\exp\left[-\frac{1}{2}\left[\left(\frac{\beta-\mu}{\sigma}\right)^2\right]\right]}{\sigma\sqrt{2\pi}} d\beta \right\} \tag{3}$$

Equation 3 simultaneously captures both the compositional statistics of the true fractions $f[a,i]$ and $f[b,i]$, and the binomial counting statistics of the point-counting measurements $n[a,i]$ and $n[b,i]$. Solving Equation 3 with the method of maximum likelihood allows us to estimate the non-zero population parameters μ and σ. Let $\hat{\mu}$ and $\hat{\sigma}$ be these maximum likelihood estimates. Then $\exp(\hat{\mu})/[\exp(\hat{\mu})+1]$ is known as the 'central value' for $f[a]$, and $\hat{\sigma}$ as the 'dispersion'. This calculation works even in the presence

ISBN 978-84-947240-2-2

of zero $n[a,i]$ or $n[b,i]$ values. It therefore solves both problems mentioned at the beginning of this abstract.

If $\sigma = 0$, then the random effects model reduces to an ordinary binomial distribution. If $n[a,i] \to \infty$ and $n[b,i] \to \infty$, it converges to a logistic normal distribution. Thus it may be argued that compositional data are a special case of the random effects model, and that the random effects model is more generally applicable than compositional data models *sensu stricto*. Equation 3 can easily be generalised from two to three or more components.

Graphically, two-component point-counting datasets can be visualised on radial plots (Galbraith, 1988). These are bivariate scatter plots that set out the normalised values of the transformed data against their precision. Therefore, radial plots are an effective visualisation tool for heteroscedastic datasets (Figure 1). Three component datasets can be displayed on ternary diagrams, like compositional data. However, for higher dimensional datasets, point-counting data and compositional data require different treatment.

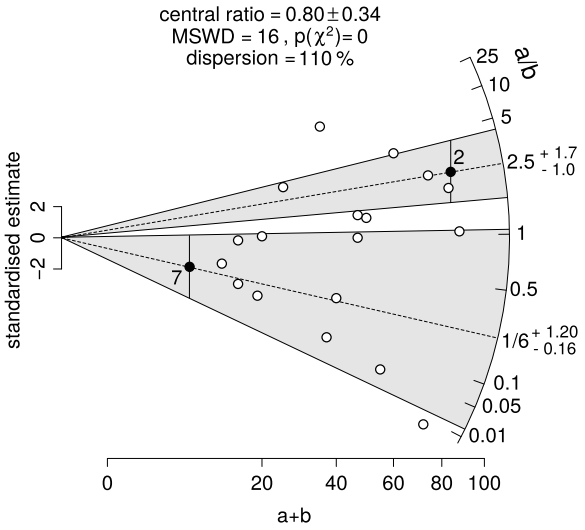


Figure 1: Radial plot of a bivariate point-counting dataset (Vermeesch, 2018). Each point on this diagram represents a single measurement of components a and b. Two of these measurements are highlighted in black. The a/b-ratio of each analysis can be read by projecting the corresponding point onto the radial scale. Its precision is obtained by projecting a 2-sigma error bar onto that same scale. Thus, the radial plot displays both the value and the precision of heteroscedastic datasets.

Principal Component Analysis (PCA) of compositional data uses Aitchison's logratio distance, which is incompatible with point-counting data due to the presence of zeros as explained before. In contrast, Correspondence Analysis (CA) is a multivariate ordination technique that is similar in purpose to PCA but uses the chi-square distance instead of the logratio distance. PCA and CA are both special cases of Multidimensional Scaling (MDS) that can be visualised as biplots.

All the methods discussed in this abstract have been been implemented in the `provenance` `R`-package, which is available from `http://provenance.london-geochron.com` (Vermeesch et al., 2016).

References

Galbraith, R. (1988). Graphical display of estimates having differing standard errors. *Technometrics 30*(3), 271–281.

Galbraith, R. F. (2005). *Statistics for fission track analysis*. CRC Press.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (2003). Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology 35*(3), 253–278.

Vermeesch, P. (2018). Statistical models for point-counting data. *Earth and Planetary Science Letters 501*, 1–7.

Vermeesch, P., A. Resentini, and E. Garzanti (2016). An R package for statistical provenance analysis. *Sedimentary Geology*.

ISBN 978-84-947240-2-2

Application of multivariate imputation of left-censored data in biomonitoring of radioisotopes

Z. Ziembik, and A. Dołhańczuk-Śródka
University of Opole, Opole, Poland; *ziembik@uni.opole.pl*

Summary

In biomonitoring living organisms, usually plants, are used in assessment of natural environment condition and in estimation of elements and chemical compounds concentration in an area.
Determination of radioisotope's activity concentration in a sample can provide valuable information. But one of the issues occurring during data interpretation is a result of low radioisotopes concentration in the material studied. Very often significant number of below MDA (Minimum Detectable Activity) communicates appear in output of signal analysis. As a result multivariate analysis of the data becomes impossible. To solve this problem data imputation method can be utilized.
In our survey samples of moss, Ol soil horizon and shallow mineral layer were collected in Bory Stobrawskie forest, located in south-western part of Poland. In the samples activity concentrations of Cs-137, K-40, Pb-210, Pb-214, Bi-214, Ac-228, Th-231, and U-235 were determined using gamma spectrometry method. The data were grouped according to the material investigated. Activity concentrations of radioisotopes were recalculated to mass fractions. In general about 30The following functions were used for data imputation: multLN, lrDA, lrEM and multRepl. The functions were used with default parameters. Usually an influence of the imputation method on the data was limited. But for a variable with numerous MDAs influence of the method on generated concentrations was significant enough to disable appropriate data imputation.

Key words: biomonitoring, radioisotope, limit of detection, imputation

1 Introduction

Biomonitoring is a method used for assessment of natural environment condition. Concentrations of elements or chemical compounds under interest are determined in living organisms. An advantage of biomonitoring over classical methods based on analysis of mineral parts of environment is better evaluation of compounds migration into trophic chains, and then better estimation of the compounds' influence on physiological state of living organisms.
Radioisotopes are common in environment. Majority of them are natural components but starting from the mid of 20. century the artificial ones also have been occurring in nature.
It seems that currently radioisotopes in environment don't pose a health risk in global scale. But actual risk may arise in a local scale, for example as a result of accidents during transport, careless storage of radioactive materials or failure in industrial installation. However, massive releases of radioactive material to environment, which have affected whole regions, also occurred. Accidental emission of radioactive materials to environment can be controlled only in a limited range. Resulting contamination on a polluted area has to be carefully monitored to apply appropriate safety measures. The key role in selection of a protection or decontamination method is recognition of transport mechanism of the radioisotopes between components of the environment, particularly migration to the living organisms.

ISBN 978-84-947240-2-2

Surveys aimed to study such mechanism involve determination of the radioisotopes activity concentration in samples of different components from the natural environment. The results are analyzed to reveal regularities in data patterns, leading to formulation of the radioisotope transport description. Activity concentration of a radioisotope can be recalculated to their mass concentrations, so the methods designed for compositional data analysis can be applied.
One of the problems occurring during experimental data analysis are observations reported as "Below Detection Limit" (BDL) or below "Limit Of Detection" (LOD). In analysis of radioisotopes spectra "Minimum Detectable Activity" (MDA) can be calculated.
The BDLs in data hinder or even makes impossible studies using methods of multidimensional data analysis. This problem can be reduced by the BDL data imputation.
Several methods of data imputations are known. The methods applied are based on different assumptions and produce different results. Particularly, the BDL limit introduces left-censored data type.

2 Results and Disscussion

Activity concentrations of gamma radioactive isotopes were determined in soil and moss *Pleurozium schreberi* samples. The material was collected in Bory Stobrawskie forest (south-western part of Poland). Activity concentration of several natural radioisotopes were determined, but also the artificial Cs-137 was found in the samples. In Tab. 1 properties of the natural radioisotopes found in samples are shown.

Table 1: The half-life, long-living ancestor and membership in decay series of the radioisotopes determined in samples

Radioisotope	Half-life	Long-living ancestor (half-life)	Decay series (origin)
K-40	$1.25{\cdot}10^{9}$a	-	- (natural)
Bi-214	19.9 min	Ra-226 (1600 a)	uranium (natural)
Pb-210	22.2 a	Ra-226 (1600 a)	uranium (natural)
Ac-228	6.15 h	Th-232 ($1.41{\cdot}10^{10}$a)	thorium (natural)
Pb-212	10.6 h	Th-232 ($1.41{\cdot}10^{10}$a)	thorium (natural)
U-235	$7.04{\cdot}10^{8}$a	own	actinium (natural)
Th-231	25.5 h	U-235 ($7.04{\cdot}10^{8}$a)	actinium (natural)
Cs-137	30.1 a	-	- (artificial)

Activity concentrations of the radioisotopes in samples were often lower than the MDA. Recalculation from activity concentration to mass fraction resulted in zero values in the data. The patterns of zero's distributions in the calculated data is illustrated in figures generated by zPatterns function from zCompositions library (Figures 1-3).

The following functions for left-censored data imputation were used: multLN, lrDA, lrEM and multRepl. To test an influence of the initial functions' settings on their practical usability in computations result their default settings remained unchanged.
As it could be expected, an effectiveness of the functions in the data imputation was different in relation to the data structure. Algorithms implemented in functions required data fulfilling specific requirements. A summary of function usability for the data imputation in moss and soil is presented in Tab. 2. In this table a function's response for a particular data structure is presented.

In relation to patterns of zero's distributions, application of the imputation functions produce more or less coherent results. For sparse concentration vectors the imputation results are useless, since they are strictly

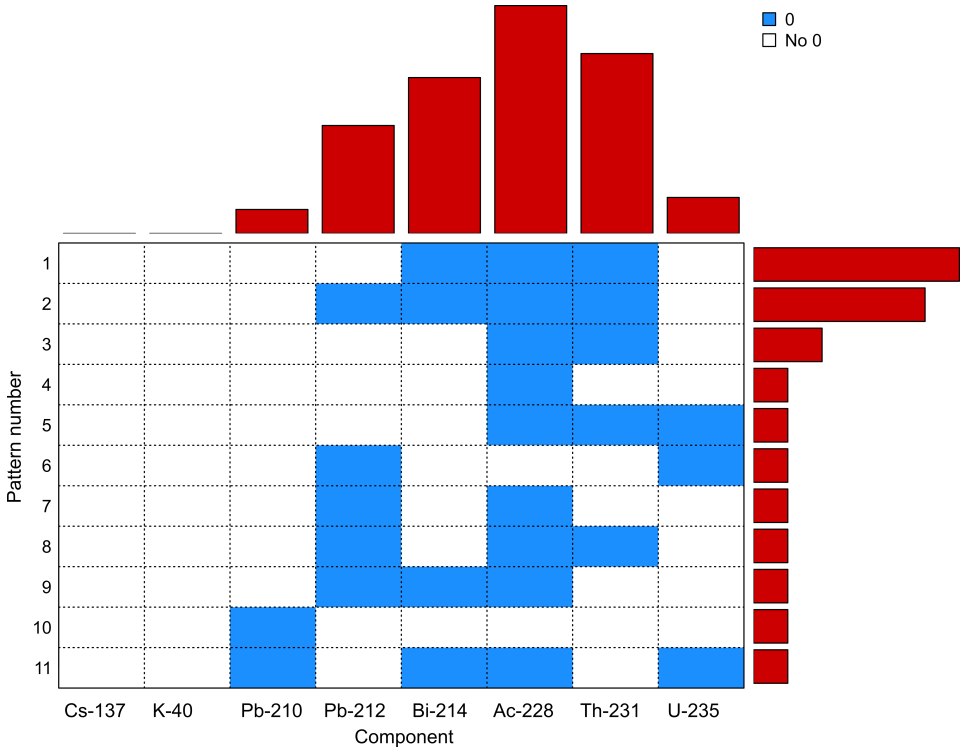


Figure 1: The structure of LODs distribution in radioisotopes concentration determined in moss.

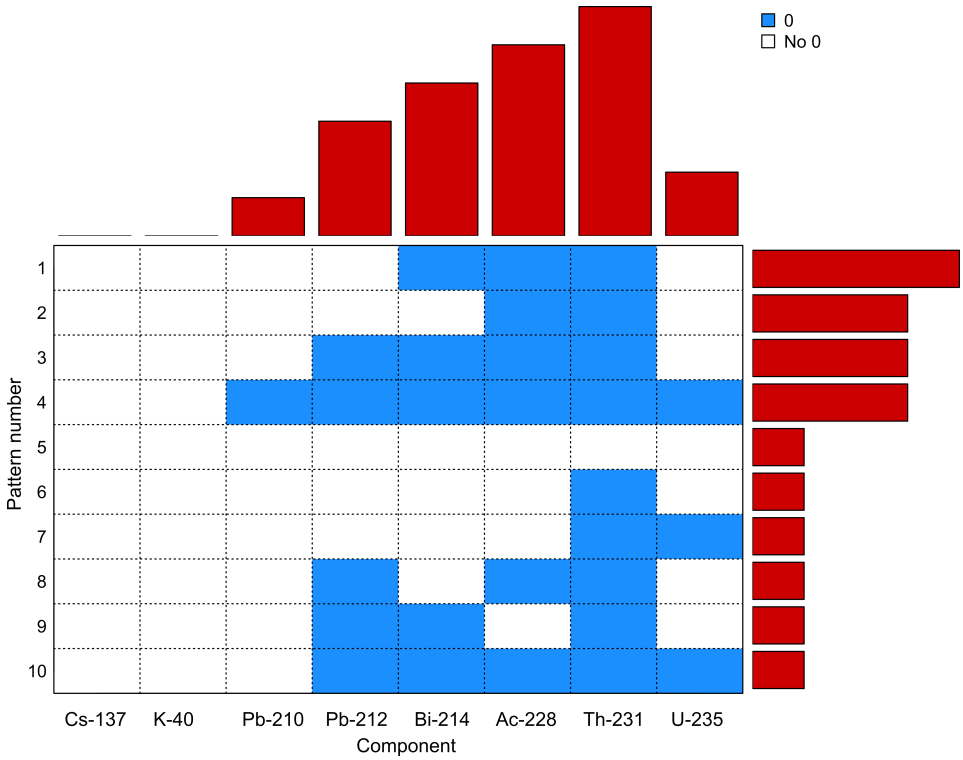


Figure 2: The structure of LODs distribution in radioisotopes concentration determined in Ol soil horizon.

 ISBN 978-84-947240-2-2

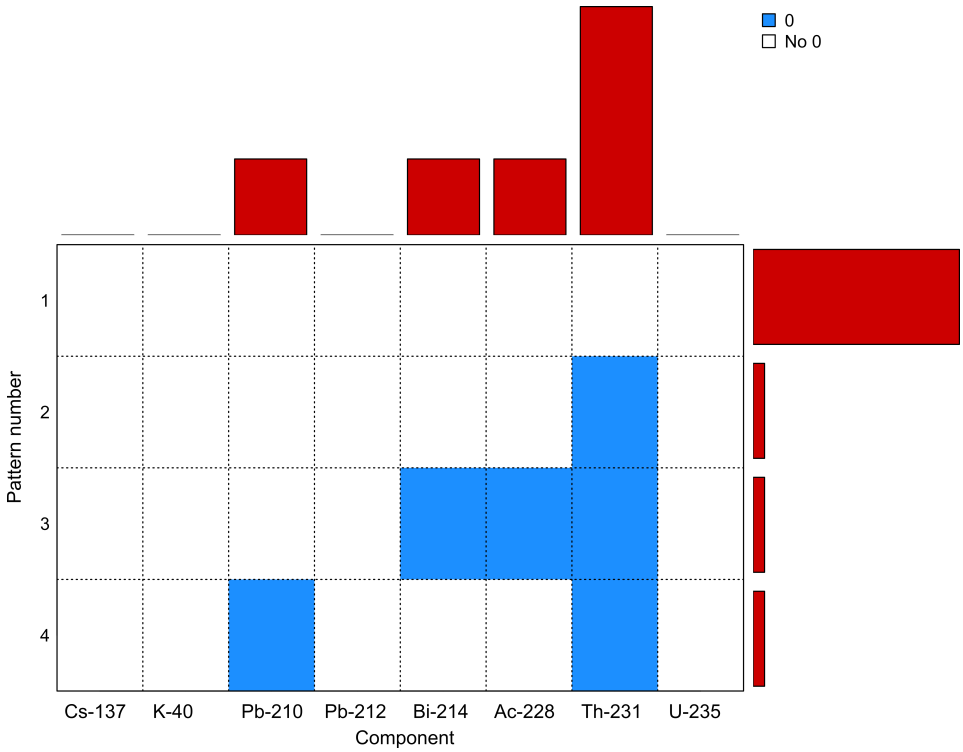


Figure 3: The structure of LODs distribution in radioisotopes concentration determined in mineral soil horizon.

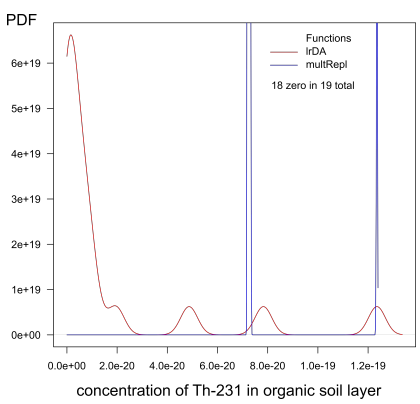
Table 2: The usabilty of the functions in data imputation. Label "err" points internal error in the function's operation, label "OK" means return of the zero imputed data frame.

	moss	Ol horizon	mineral layer
multLN	OK	err	OK
lrDA	err	OK	OK
lrEM	err	err	OK
multRepl	OK	OK	OK

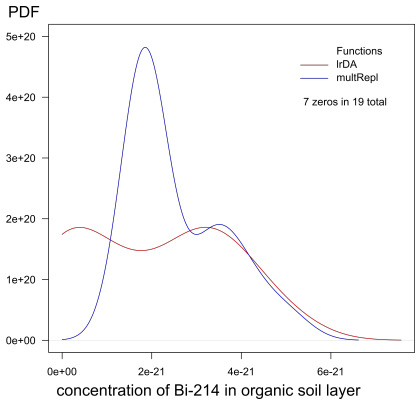
related to the function used for calculations. The imputation functions can produce artificial data even if actually no information about a variable distribution is provided. An example of the issue is illustrated in Fig. 4A. Only one concentration of Th-231 in organic soil layer was bigger than MDA. For lrDA and multRepl functions this information was enough to compute the desired data. But the results are dissimilar and they useless in computations. A problem in interpretation of data analysis result arises when one uncritically applies such data in computations.

In Fig. 4B distribution of Bi-214 concentration in organic soil layers is illustrated. Among 19 results 7 of them were below MDA. For low Bi-214 concentration, the 37% of imputed data produce an incoherence in imputation results. It is not clear which function is better for calculations. A solution of the issues may be related to optimal tuning of the functions' parameters. In Fig. 4C distribution of Bi-214 in mineral soil layer is shown. Among 21 concentrations only 1 was imputed. But this single insertion produces somewhat different distributions of the variable imputed by different functions.

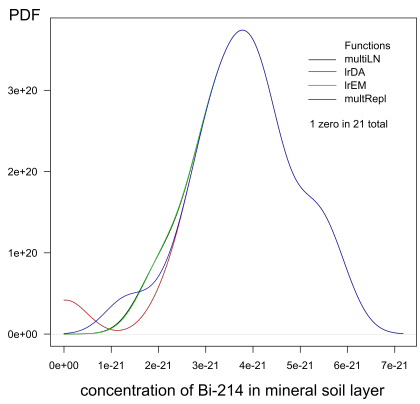
 ISBN 978-84-947240-2-2



(A) Th-231 concentration in organic soil layer



(B) Bi-214 concentration in organic soil layer



(C) Bi-214 concentration in mineral soil layer

Figure 4: Examples of concentration distributions (PDF) for various number of zeros in the data

3 Conclusions

The R library zCompositions offers a number of very well developed imputation methods but they should not be applied in a random style. A method of "by chance" application of imputation functions doesn't deliver reliable approach to the data analysis. Selection of imputation function is not obvious, it should be preceded by analysis of the nature of the phenomenon studied and methods applied in the investigated system evaluation. Comparison of functions' outputs is recommended prior to actual data analysis.

References

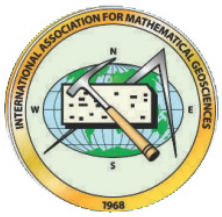
R Development Core Team. (2019). R: A language and environment for statistical computing. R foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org..

Palarea-Albaladejo, J., and Martín-Fernández, J. A. (2015). zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems 143*, pp. 85–96.

Palarea-Albaladejo, J., and Martín-Fernández, J. A. (2015). zCompositions: Imputation of Zeros and Nondetects in Compositional Data Sets. R package version 1.2.0. http://CRAN.R-project.org/package=zCompositions.

CoDaWork2019 has been sponsored by:

Ajuntament de Terrassa



International Association of Mathematical Geosciences

Statistical Modelling Society (SMS)

Societat Catalana d'Estadística (SoCE)

Journal SORT

Universitat Politècnica de Catalunya-BarcelonaTECH

ISBN: 978-84-947240-2-2