




Classical and Robust Regression Analysis with Compositional Data

K. G. van den Boogaart¹ · P. Filzmoser² ·
K. Hron³ · M. Templ⁴  · R. Tolosana-Delgado¹

Received: 15 November 2019 / Accepted: 14 September 2020
© The Author(s) 2020

Abstract Compositional data carry their relevant information in the relationships (logratios) between the compositional parts. It is shown how this source of information can be used in regression modeling, where the composition could either form the response, or the explanatory part, or even both. An essential step to set up a regression model is the way how the composition(s) enter the model. Here, balance coordinates will be constructed that support an interpretation of the regression coefficients and allow for testing hypotheses of subcompositional independence. Both classical least-squares regression and robust MM regression are treated, and they are compared within different regression models at a real data set from a geochemical mapping project.

✉ K. G. van den Boogaart
boogaart@math.tu-freiberg.de

P. Filzmoser
p.filzmoser@tuwien.ac.at

K. Hron
karel.hron@upol.cz

M. Templ
matthias.templ@zhaw.ch

R. Tolosana-Delgado
r.tolosana@hzdr.de

- ¹ Helmholtz Institut Freiberg for Resources Technology, Freiberg, Germany
- ² Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Vienna, Austria
- ³ Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, Olomouc, Czech Republic
- ⁴ Institute for Data Analysis and Process Design, Zurich University of Applied Sciences, Winterthur, Switzerland

Keywords Balances · Robust regression · GEMAS project · Hypothesis testing · Robust bootstrap

1 Introduction

Although regression analysis belongs to the most developed statistical procedures, there is not too much literature available if compositional data are involved (see, e.g., Aitchison 1986; Daunis-i-Estadella et al. 2002; Tolosana-Delgado and van den Boogaart 2011; van den Boogaart and Tolosana-Delgado 2013; Egozcue et al. 2013; Pawlowsky-Glahn et al. 2015; Fišerová et al. 2016; Coenders et al. 2017; Filzmoser et al. 2018; Greenacre 2019). Regression with compositional data presents certain particularities because of the special statistical scale of compositions. Compositions have been typically (and restrictively) defined as vectors of positive components summing up to a constant, with the simplex as their sampling space (Aitchison 1986). However, since the beginning of the twenty-first century, it has become clearer that data may not abide to the constant sum condition and nevertheless be sensibly considered as compositional, the determining point being the relativity of the information provided (Aitchison 1997; Barceló-Vidal et al. 2001). That means, when dealing with compositional data the crucial quantities are formed by the ratios between the compositional parts (i.e., the variables) rather than by the reported data values directly. There are different proposals to extract this relative information, frequently referred to as transformations or coordinates, with the additive logratio (alr), the centered logratio (clr), and the isometric logratio (ilr) transformation as the most well-known representatives (van den Boogaart and Tolosana-Delgado 2013; Pawlowsky-Glahn et al. 2015; Filzmoser et al. 2018). The scores obtained with such transformations can be referred to as “coordinates” thanks to the Euclidean space structure of the simplex (Billheimer et al. 2001; Pawlowsky-Glahn and Egozcue 2001a), described later on in Sect. 2.

Due to its affine equivariance, linear regression is known to provide exactly the same result whichever logratio transformation is used (van den Boogaart and Tolosana-Delgado 2013), although the use of the clr transformation may entail unnecessary numerical complications (the generalised inversion of a singular covariance matrix) whenever the composition plays an explanatory role. Even more, regression models with compositional data can be established without using any logratio transformation at all. Indeed, this fact allows to consider the objects occurring in the regression analysis (slopes, intercepts, gradients, residuals, predictions, etc) as fully intrinsically compositional objects, the use of one or another logratio transformation being a choice of representation. This paper just works with an isometric logratio representation because of its intimate connection with tests for exclusion of single components or subcompositions.

Generally, one must distinguish different types of regression models involving compositional data. The composition could form the response part in the regression model, and the explanatory part is consisting of one or more non-compositional features. This leads to a multivariate regression model, later on denoted as Type 1 model. The reverse problem, composition as explanatory part, non-compositional response, is denoted as Type 2 model in this contribution; if the response is only univariate, we end up with a

multiple linear regression model. The Type 3 model is characterized by a composition as explanatory part and another composition as response part, thus becoming a multivariate multiple linear regression model. For each considered type, we can construct the regression model exploiting the Euclidean structure of the simplex, or else in terms of ilr coordinates or clr transformed scores. However, since there are infinitely many possibilities to construct ilr coordinates (Egozcue et al. 2003), one can focus on those coordinates that allow for an interpretation of the model and the corresponding regression coefficients. Even more, if particular hypotheses shall be tested, it is essential to construct such coordinates that support the testing problem. This contribution focuses particularly in so called tests of subcompositional independence, or rather uncorrelation. Here the goal is to establish subcompositions (or single components) that do not depend on or do not influence the covariables considered.

A further issue which is also relevant for traditional non-compositional regression is parameter estimation. The most widely used method is least-squares regression, minimizing the sum of squared residuals. In presence of outliers in the explanatory or/and response variable(s) it is known that this estimator can be heavily biased, and robust counterparts should be preferred (Maronna et al. 2006). Here, the highly robust MM estimator for regression is considered, for both the multiple and the multivariate linear regression model (Van Aelst and Willems 2013). There is also technique available for robust statistical inference (Salibián-Barrera et al. 2008). Such techniques are of the utmost importance in compositional data, because here small values (near or below determination limits, rounded zeros, counting zeros) can otherwise become highly influential after the logratio transformations.

In this contribution we thoroughly define the different types of regression models involving compositional data, with geometric interpretations, and explain their use in a case study. In Sect. 2 we provide more detailed information about compositional data analysis, and define the three considered types of models. Section 3 refers to least-squares estimation, and depending on the model it is shown how parameter estimation and hypothesis testing can be carried out. Estimation and hypothesis testing using robust MM regression is treated in Sect. 4. All three linear models and classical as well as robust estimation are illustrated in Sect. 5 with a data set originating from the GEMAS project, a continental scale soil geochemistry survey in Europe. The final Sect. 6 summarizes and concludes.

2 Compositional Data Analysis

2.1 Compositional Geometry

A (column-)vector $\mathbf{x} = [x_1, x_2, \dots, x_D]$ is considered a D -part composition if its components do inform on the relative importance of a set of parts forming a total. Because of this, compositions are often identified with vectors of positive components and constant or irrelevant sum, i.e. vectors that can be reclosed to sum up to 1 (or to any other constant) by $\mathcal{C}[\mathbf{x}] = (\mathbf{1}' \cdot \mathbf{x})^{-1} \mathbf{x}$, without loss of relevant information. In the eighties of the twentieth century Aitchison (1986) proposed to statistically treat this kind of data after some sort of one-to-one multivariate logratio transformation,

generalising the logit, and he defined several of them, most notably the centered logratio (clr) transformation (Aitchison 1986)

$$\text{clr}(\mathbf{x}) = \ln \frac{\mathbf{x}}{\sqrt[D]{\prod x_i}}, \quad \left(\text{inverse: } \text{clr}^{-1}(\mathbf{x}^*) = \mathcal{C}[\exp(\mathbf{x}^*)] \right),$$

with logs and exponentials applied component-wise. The sample space of a random composition, the simplex \mathcal{S}^D , has been since the end of the nineties recognized to have a vector space structure, induced by the operations of perturbation $\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, x_2 y_2, \dots, x_D y_D]$ (Aitchison 1986) and powering $\lambda \odot \mathbf{x} = \mathcal{C}[x_1^\lambda, x_2^\lambda, \dots, x_D^\lambda]$ (Aitchison 1997). The neutral element with respect to this structure is proportional to a vector of D ones, $\mathbf{n} = \mathcal{C}[\mathbf{1}_D]$. This structure can be extended (Pawlowsky-Glahn and Egozcue 2001b) to a Euclidean space structure (Aitchison et al. 2002) by the scalar product

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} = \langle \text{clr}(\mathbf{x}), \text{clr}(\mathbf{y}) \rangle, \tag{1}$$

which induces a compositional distance (Aitchison 1986)

$$d_A^2(\mathbf{x}, \mathbf{y}) = \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \left[\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right]^2, \tag{2}$$

both fully compliant with the concept of relative importance conveyed in the modern definition of compositions.

To take this relative character into account in an easy fashion when statistically analyzing compositional data sets, the principle of working on coordinates is recommended (Pawlowsky-Glahn 2003). This states that the statistical analysis should be applied to the coordinates of the compositional observations, preferably in an orthonormal basis of the Euclidean structure $\{\mathcal{S}^D, \oplus, \odot, \langle \cdot, \cdot \rangle_A\}$, and that results might be expressed back as compositions by applying them to the basis used. A simple and easy way to compute orthonormal coordinates is provided by the isometric log-ratio (ilr) transformation (Egozcue et al. 2003)

$$\text{ilr}(\mathbf{x}) =: \mathbf{x}^* = \mathbf{V}^t \cdot \ln \mathbf{x} = \mathbf{V}^t \cdot \text{clr}(\mathbf{x}), \tag{3}$$

where $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_{D-1})$ is a matrix with orthonormal columns, each orthogonal to the vector of ones, i.e.

$$\mathbf{v}_i^t \cdot \mathbf{v}_j = \delta_{ij} \quad \text{and} \quad \mathbf{1}_D^t \cdot \mathbf{v}_i = 0. \tag{4}$$

Each of these columns provide the clr coefficients of each of the compositional vectors forming the orthonormal basis used, so that the orthonormal basis on the simplex (with respect to the Aitchison geometry) is $\mathbf{w}_i = \text{clr}^{-1}(\mathbf{v}_i)$. Conversely, given a vector of

coordinates $\mathbf{x}^* = [x_1^*, x_2^*, \dots, x_{D-1}^*]$, the inverse isometric transformation provides a convenient way to apply it to the basis

$$\bigoplus_{i=1}^{D-1} x_i^* \odot \mathbf{w}_i =: \text{ilr}^{-1}(\mathbf{x}^*) = \mathcal{C}[\exp(\mathbf{V} \cdot \mathbf{x}^*)], \tag{5}$$

and the clr coefficients can be recovered as well from the ilr coordinates with

$$\text{clr}(\mathbf{x}) = \mathbf{V} \cdot \text{ilr}(\mathbf{x}). \tag{6}$$

There are infinitely many ilr transformations, actually as many as matrices satisfying the conditions of Eq. (4). Each represents a rotation of the coordinate system (see, e.g., Egozcue et al. 2003; Filzmoser et al. 2018). For the purposes of this contribution, it is relevant to consider those matrices linked to particular subcompositions, i.e. subsets of components. If one wants to split the parts in \mathbf{x} into two groups, say the first s against the last $r = D - s$, there is a vector that identifies the balance between the two groups,

$$\mathbf{v} = \frac{1}{\sqrt{rs^2 + sr^2}} \left[\underbrace{r, r, \dots, r}_s, \underbrace{-s, -s, \dots, -s}_r \right]. \tag{7}$$

Given that one can always split the resulting two subcompositions again into two groups, this sort of structure can be reproduced $D - 1$ times, generating $D - 1$ vectors of this kind with only three values every time (a positive value for one group, a negative value for the other group and zero for those parts not involved in that particular splitting). This is called a sequential binary partition, and the interested reader may find the details of the procedure in Egozcue and Pawlowsky-Glahn (2005, 2011).

A particular case occurs when the splitting of interest places one individual variable against all other (a sort of “one component subcomposition”). Considering the first component as the one that is desired to be isolated, the balancing vector is then obtained with Eq. (7) taking $r = 1$ and $s = D - 1$. Balances isolating any other part can be obtained permuting the resulting components. This sort of balances, which can be identified with so-called pivot coordinates (Filzmoser et al. 2018), are thus useful also to check the possibility to eliminate one single part from the regression model.

2.2 Compositional Linear Models

Three kinds of linear models involving compositions have been defined (van den Boogaart and Tolosana-Delgado 2013; Pawlowsky-Glahn et al. 2015; Filzmoser et al. 2018): models with compositional response (Aitchison 1986; Daunis-i-Estadella et al. 2002), models with compositional explanatory variable (Aitchison 1986; Tolosana-Delgado and van den Boogaart 2011), and models with compositions as both explanatory variable and response (Egozcue et al. 2013). The next sections systematically build each regression model solely by means of geometric operations, and show

how these models are then represented in an arbitrary isometric logratio transformation.

2.2.1 Linear Model with Compositional Response (Type 1)

A model with compositional response assumes that a random composition \mathbf{Y} is a linear function (in the sense of the Aitchison geometry) of several explanatory real random variables X_0, X_1, \dots, X_P , which gives the expected value of some normally distributed composition,

$$\hat{\mathbf{Y}} = \bigoplus_{i=0}^P X_i \odot \mathbf{b}_i, \quad \mathbf{Y} \sim \mathcal{N}_{\mathcal{G}^D}(\hat{\mathbf{Y}}, \boldsymbol{\Sigma}_\varepsilon), \quad (8)$$

where $\mathcal{N}_{\mathcal{G}^D}(\hat{\mathbf{Y}}, \boldsymbol{\Sigma}_\varepsilon)$ stands for the normal distribution on the simplex of \mathbf{Y} (Mateu-Figueras and Pawłowsky-Glahn 2008), parametrized in terms of a compositional mean vector and a covariance matrix of the random composition in some ilr representation. This reflects the fact that the normal distribution on the simplex of a random composition corresponds to the (usual) normal distribution of its ilr representation. This regression model is useful for explanatory variables of type quantitative (regression), categorical (ANOVA) or a combination of both (ANCOVA). Note that one can establish this regression model for compositional data in a least-square sense (Mood et al. 1974, Chapter X), free of the normality assumption, by using the Aitchison distance [Eq. (2)] as Daunis-i-Estadella et al. (2002) proposed. However, the normality assumption is needed in the context of hypotheses testing which is one of the main contributions of this paper. Specifically, it serves for deriving the distribution of the test statistics in the classical (least squares) regression case and serves also as the reference model for robust regression.

If a logratio transformation is applied to this model, this yields a conventional multiple, multivariate linear regression model on coordinates

$$\hat{\mathbf{Y}}^* = \sum_{i=0}^P X_i \cdot \mathbf{b}_i^*, \quad \mathbf{Y}^* \sim \mathcal{N}^{D-1}(\hat{\mathbf{Y}}^*, \boldsymbol{\Sigma}_\varepsilon). \quad (9)$$

The model parameters are thus the slopes $\mathbf{b}_0^*, \mathbf{b}_1^*, \dots, \mathbf{b}_P^*$, and the residual covariance matrix $\boldsymbol{\Sigma}_\varepsilon$. Note that it is common to take $X_0 \equiv 1$ and then \mathbf{b}_0^* rather represents the intercept of the model in the logratio coordinate system chosen. The specification in Eq. (9) has the advantage to be tractable with conventional software and solving methods. Once estimates of the vector coefficients are available, they can be back-transformed to compositional coefficients, e.g. $\hat{\mathbf{b}}_i = \text{ilr}^{-1}(\hat{\mathbf{b}}_i^*)$ if calculations are done in ilr coordinates. Alternatively, ilr coordinates can also be converted to clr coefficients with $\hat{\mathbf{b}}_i^{\text{clr}} = \mathbf{V} \cdot \hat{\mathbf{b}}_i^*$.

It is important to emphasise that the predictions provided by this regression model are unbiased both in terms of any logratio representation [Eq. (9)], and in terms of the original composition [Eq. (8)] with respect to the Aitchison geometry discussed in

Sect. 2.1. This follows directly from the isometry of the ilr or clr mappings (Egozcue et al. 2012; Pawlowsky-Glahn et al. 2015; Fišerová et al. 2016). If interest lies on understanding the unbiasedness properties of predictions (8) with respect to the conventional Euclidean geometry of the real multivariate space \mathbb{R}^D , i.e. on the nature of the expected value of $\hat{\mathbf{Y}} - \mathbf{Y}$, then one can resort to numerical integration of the model explicated by Eq. (8), which provides the conditional distribution of \mathbf{Y} given $\hat{\mathbf{Y}}$ (Aitchison 1986).

2.2.2 Regression with Compositional Explanatory Variable (Type 2)

A model with a compositional explanatory variable \mathbf{X} and one explained real variable Y is (both in composition and coordinates)

$$\hat{Y} = b_0 + \langle \mathbf{b}, \mathbf{X} \rangle_A = b_0 + [\mathbf{b}^*]^t \cdot \mathbf{X}^* = b_0 + \langle \text{clr}(\mathbf{b}), \text{clr}(\mathbf{X}) \rangle, \quad Y \sim \mathcal{N}(\hat{Y}, \sigma_\varepsilon^2).$$

The model parameters are thus the intercept b_0 , the gradient \mathbf{b}^* and the residual variance σ_ε^2 , which again can be estimated with any conventional statistical toolbox. The gradient, once estimated in coordinates, can be back-transformed to a compositional gradient as $\hat{\mathbf{b}} = \text{ilr}^{-1}(\mathbf{b}^*)$, or to its clr representation by $\hat{\mathbf{b}}_i^{\text{clr}} = \mathbf{V} \cdot \hat{\mathbf{b}}_i^*$. Note that solving a Type 2 model directly in clr would require the use of generalised inversion of the covariance matrix of $\text{clr}(\mathbf{X})$, which provides the same results but at a higher computational cost.

2.2.3 Compositional to Compositional Regression (Type 3)

A model with both an explanatory $\mathbf{X} \in \mathcal{S}^{D_x}$ and an explained $\mathbf{Y} \in \mathcal{S}^{D_y}$ compositional variables can be expressed in several ways. This time, the easiest is directly in ilr coordinates,

$$\hat{\mathbf{Y}}^* = \mathbf{b}_0^* + \mathbf{B}^* \cdot \mathbf{X}^*, \quad \mathbf{Y}^* \sim \mathcal{N}^{D-1}(\hat{\mathbf{Y}}^*, \boldsymbol{\Sigma}_\varepsilon), \tag{10}$$

with model parameters a $(D_y - 1)$ -component vector intercept \mathbf{b}_0^* , a $(D_y - 1) \times (D_y - 1)$ -element residual covariance matrix $\boldsymbol{\Sigma}_\varepsilon$, and a $(D_y - 1) \times (D_x - 1)$ -matrix of slope coefficients $\mathbf{B}^* = [b_{ij}^*]$. Note that in this composition-to-composition model it is necessary to distinguish between the number of components D_x versus D_y and the transformations ilr_x and ilr_y used for each of the two compositions. In this representation, again, the model parameters can be estimated with any available tool, and then we can interpret them in compositional terms. The intercept $\mathbf{b}_0 = \text{ilr}_y^{-1}(\mathbf{b}_0^*)$ is the expected response composition when the explanatory composition has a neutral value $\mathbf{X} = \mathbf{n}_x$. The covariance matrix $\boldsymbol{\Sigma}_\varepsilon$ has the same interpretation as in the model with compositional response (Eq. 9), as the variance of the response conditional on the explanatory variable. Finally, the slope matrix can be seen in three different ways. First, each row represents the gradient $\mathbf{b}_i = \text{ilr}_x^{-1}[b_{i1}^*, b_{i2}^*, \dots, b_{i(D_x-1)}^*] \in \mathcal{S}^{D_x}$ along which one particular ilr response coordinate increases fastest,

$$\hat{Y}_i^* = (b_0^*)_i + \langle \mathbf{b}_i, \mathbf{X} \rangle_A = (b_0^*)_i + [b_{i1}^*, b_{i2}^*, \dots, b_{i(D_x-1)}^*]^t \cdot \mathbf{X}^*.$$

Second, each column represents the slope $\mathbf{b}_i = \text{ilr}_y^{-1}[b_{1i}^*, b_{2i}^*, \dots, b_{(D_y-1)i}^*] \in \mathcal{S}^{D_y}$ along which each explanatory ilr coordinate can modify the response,

$$\hat{\mathbf{Y}}^* = \mathbf{b}_0 \oplus \bigoplus_{i=1}^P X_i^* \odot \mathbf{b}_i.$$

Third, one can interpret \mathbf{B}^* as the matrix representation (on the chosen bases of the two simplexes) of a linear application $\mathbf{B}: S^{D_x} \rightarrow S^{D_y}$, which is nothing else than the combination of a rotation on S^{D_x} and a rotation on S^{D_y} , together with a univariate linear regression of each of the pairs of rotated axes:

$$\mathbf{B}^* = \sum_{i=1}^R \mathbf{u}_i \cdot d_i \cdot \mathbf{v}_i^t, \quad R \leq \min(D_x, D_y). \tag{11}$$

Here, the matrix $\mathbf{U} = [\mathbf{u}_i]$ of left vectors is the rotation on the image simplex S^{D_y} , and that of the right vectors $\mathbf{V} = [\mathbf{v}_i]$ the rotation on the origin simplex S^{D_x} . The coefficient d_i is then the slope of the regression between the pair of rotated directions. Note that this representation coincides with a singular value decomposition of the matrix \mathbf{B}^* , and is reminiscent of methods such as canonical correlation analysis or redundancy analysis (Graffelman and van Eeuwijk 2005). To recover clr representations of the matrix of coefficients, or of these singular vectors, one just needs the respective basis matrices \mathbf{V}_x and \mathbf{V}_y ,

$$\mathbf{B}^{\text{clr}} = \mathbf{V}_y \cdot \mathbf{B}^* \cdot \mathbf{V}_x^t, \quad \mathbf{u}_i^{\text{clr}} = \mathbf{V}_x \cdot \mathbf{u}_i^*, \quad \mathbf{v}_i^{\text{clr}} = \mathbf{V}_y \cdot \mathbf{v}_i^* \tag{12}$$

These expressions apply to the model coefficients (\mathbf{B}^*) and to their estimates ($\hat{\mathbf{B}}^*$) given later on in Sects. 3 and 4.

Note that the same issues about the unbiasedness of predictions raised in Sect. 2.2.1 apply to predictions obtained with Eq. (10).

2.3 Subcompositional Independence

One of the most common tasks of regression is the validation of a particular model against data, in particular models of (linear) independence, partial or complete. In a non-compositional framework, independence is identified with a slope or gradient matrix/vector identically null (complete independence), or just with some null coefficients (partial independence). Complete independence for compositional models is also identified with a null slope, null gradient vectors, or null matrices of the model established for coordinates (each slope \mathbf{b}_i^* , the gradient \mathbf{b}^* resp. the matrix \mathbf{B}^*). But partially nullifying one single coefficient of these vectors or matrices just forces independence of the covariable(s) with a certain logratio, not with the components this logratio involves. The necessary concept in this context is thus rather one of subcompositional independence, i.e. that a whole subset of components has not influence in

resp. is not influenced by a covariable. One must further distinguish two cases, namely internal and external subcompositional independence.

Consider the first s components of the composition as independent of a given covariable. One can then construct a basis of \mathcal{S}^D with three blocks: an arbitrary basis of $s - 1$ vectors comparing the first s components (independent subcomposition), the balancing vector between the two subcompositions (Eq. 7), and an arbitrary basis of $r - 1$ vectors comparing the last $r = D - 1$ components (dependent subcomposition).

In a Type 1 regression model (compositional response), internal independence of a certain subcomposition with respect to the i th explanatory covariable X_i means that this covariable is unable to modify the relations between the components of the independent subcomposition, i.e. $b_{1i}^* = b_{2i}^* = \dots = b_{(s-1)i}^* = 0$. External independence further assumes that the balance coordinate is independent of the covariable, $b_{si}^* = 0$.

In a Type 2 regression model (compositional input), internal independence means that the explained covariable Y cannot change due to variations within the independent subcomposition, i.e. that the coordinate gradient satisfies $b_1^* = b_2^* = \dots = b_{(s-1)}^* = 0$. External independence further assumes that the explained variable only depends on the relationships within the dependent subcomposition, thus additionally $b_s^* = 0$.

Subcompositional independence for a Type 3 regression model (composition-to-composition) inherits from the concepts mentioned before. The response subcomposition formed by its first s_y parts is internally independent of the input subcomposition formed by its first s_x parts if no modification within the latter can induce any change within the former, i.e. $b_{ij}^* = 0$ for $i = 1, 2, \dots, (s_y - 1)$ and $j = 1, 2, \dots, (s_x - 1)$. Similarly, external independence further assumes $b_{ij}^* = 0$ for $i = 1, 2, \dots, s_y$ and $j = 1, 2, \dots, s_x$, including regression coefficients involving the balancing elements as well.

If the case study or question at hand does not suggest a subcomposition or subcompositions to test for such subcompositional independence hypotheses, candidates can be found with the following heuristic, inspired by the Q-mode clustering of compositional parts (van den Boogaart and Tolosana-Delgado 2013; Filzmoser et al. 2018). For a certain vector of regression coefficients \mathbf{b}_i^* (a slope, a gradient, a row or a column of a Type 3 coefficient matrix \mathbf{B}^*), one can obtain the vector of clr coefficients by $\mathbf{b}_i^{\text{clr}} = \mathbf{V} \cdot \mathbf{b}_i^*$. The key idea now is to realize that if $b_{ij}^{\text{clr}} - b_{ik}^{\text{clr}} = 0$ then $\log(X_j/X_k)$ does not influence Y_i , resp. X_i does not influence $\log(Y_j/Y_k)$. Hence we can compute the matrix of interdistances between the clr variables $d_{jk}^2 = (\text{clr}_j(\mathbf{b}_i) - \text{clr}_k(\mathbf{b}_i))^2$, and apply any hierarchical clustering technique, naturally defining an ilr basis that isolates those subcompositions which clr coefficients are most similar, and in consequence more probably are not influenced by resp. do not influence the i th covariable.

3 Classical Least Squares (LS) Estimation

3.1 LS Estimation in Regression Type 2

We denote the n observations of the response by y_1, \dots, y_n , and those of the explanatory logratio-transformed compositions by the vectors $\mathbf{x}_1^*, \dots, \mathbf{x}_n^*$ of D components (i.e., $D - 1$ logratio coordinates plus a one in the first component to account for the

intercept, if that is included in the model). The regression parameters are denoted by the vector $\mathbf{b}^* = [b_0^*, b_1^*, \dots, b_{D-1}^*]^t$, and the scale of the residuals is σ_ε . The residuals are denoted as $r_i(\mathbf{b}^*) = y_i - [\mathbf{b}^*]^t \mathbf{x}_i^*$, for $i = 1, 2, \dots, n$.

Considering the vector of all responses $\mathbf{y} = [y_1, \dots, y_n]^t$ and the matrix of all explanatory variables $\mathbf{X}^* = [\mathbf{x}_1^*, \dots, \mathbf{x}_n^*]$ (each row is an individual, the first column is the constant 1, and each subsequent column an ilr coordinate), the least squares estimators of the model parameters are

$$\widehat{\mathbf{b}}^* = [(\mathbf{X}^*)^t \cdot \mathbf{X}^*]^{-1} \cdot (\mathbf{X}^*)^t \cdot \mathbf{y}$$

and

$$\hat{\sigma}_\varepsilon^2 = \frac{1}{N - D} \sum_{i=1}^n r_i^2(\widehat{\mathbf{b}}^*).$$

Finally, the covariance matrix of $\widehat{\mathbf{b}}^*$ can be estimated as

$$\widehat{\Sigma}_{\mathbf{b}} = \hat{\sigma}_\varepsilon^2 \cdot [(\mathbf{X}^*)^t \cdot \mathbf{X}^*]^{-1}.$$

3.2 LS Estimation in Regression Types 1 and 3

When the response is compositional, we obtain observed logratio score vectors $\mathbf{y}_1^*, \dots, \mathbf{y}_n^*$ of dimension $D - 1$, and the regression coefficients are collected in the $(D - 1) \times (P + 1)$ matrix \mathbf{B}^* . The first column of this matrix represents the intercept coordinate vector \mathbf{b}_0^* . The remaining columns can be linked to P explanatory real covariables (Type 1 regression) or to the $P = D_x - 1$ ilr-transformed explanatory composition (Type 3 regression). The residual vectors are $\mathbf{r}_i^*(\mathbf{B}^*) = \mathbf{y}_i^* - \mathbf{B}^* \mathbf{x}_i$, for $i = 1, \dots, n$.

Considering the matrices of explanatory and response variables $\mathbf{X}^* = [\mathbf{x}_1^*, \dots, \mathbf{x}_n^*]$ and $\mathbf{Y}^* = [\mathbf{y}_1^*, \dots, \mathbf{y}_n^*]$ (each row is an individual, compositions are ilr-transformed), the least squares estimators of the model parameters are

$$\widehat{\mathbf{B}}^* = [(\mathbf{X}^*)^t \cdot \mathbf{X}^*]^{-1} \cdot (\mathbf{X}^*)^t \cdot \mathbf{Y}^*$$

and

$$\widehat{\Sigma}_\varepsilon = \frac{1}{N - P} \sum_{i=1}^n \mathbf{r}_i^*(\mathbf{B}^*)^t \cdot \mathbf{r}_i^*(\mathbf{B}^*).$$

Finally, the covariance matrix of $\widehat{\mathbf{b}}^*$ can be estimated as

$$\widehat{\Sigma}_{\mathbf{b}} = \widehat{\Sigma}_\varepsilon \otimes [(\mathbf{X}^*)^t \cdot \mathbf{X}^*]^{-1},$$

where \otimes is the Kronecker product of the two matrices, and $\widehat{\mathbf{B}}^*$ is vectorized stacked by columns.

3.3 LS Testing of the Subcompositional Independence

The classical theory of linear regression modeling provides a wide range of tests on regression parameters, both in univariate regression (Type 2) and multivariate regression models (Types 1 and 3) (Johnson and Wichern 2007). Among them, we are particularly interested in those that are able to cope with subcompositional independence (in its external and internal forms, respectively), as introduced in Sect. 2.3. For the model of Type 2 and the internal subcompositional independence, the corresponding hypothesis on the regression parameters can be expressed as $\mathbf{A}\mathbf{b}^* = \mathbf{0}$ with $\mathbf{A} = (\mathbf{0}, \mathbf{I})$, where \mathbf{I} is the $(s - 1) \times (s - 1)$ identity matrix and $\mathbf{0}$ stands for an $(s - 1) \times (D - s + 1)$ matrix with all its elements zero. In the alternative hypothesis the former equality does not hold. Note that for the case of external subcompositional independence, the sizes of the matrices \mathbf{I} and $\mathbf{0}$ would just change to $s \times s$ and $s \times (D - s)$, respectively. In the following, only the internal subcompositional independence will be considered, the case of the external independence could be derived analogously. Under the model assumptions including normality on the simplex and the above null hypothesis, the test statistic

$$T = \frac{(S_R - S)/(s - 1)}{S/(n - D)}, \tag{13}$$

where

$$S = \sum_{i=1}^n r_i(\widehat{\mathbf{b}}^*), \quad S_R = \sum_{i=1}^n r_i(\widehat{\mathbf{b}}_R^*),$$

follows an F distribution with $s - 1$ and $n - D$ degrees of freedom. Here, $\widehat{\mathbf{b}}_R^*$ denotes the LS estimates under the null hypothesis (i.e. just the submodel is taken for the estimation of regression parameters). The hypothesis on internal subcompositional independence is rejected if $t \geq F_{s-1, n-D}(1 - \alpha)$, the $(1 - \alpha)$ -quantile of that distribution. This test statistic coincides with the likelihood ratio test on the same hypothesis, that can be easily generalized for the case of multivariate regression. The statistic can be written also in the form

$$T = \frac{S_R^* - S^*}{s - 1}$$

for

$$S^* = \sum_{i=1}^n \left(\frac{r_i(\widehat{\mathbf{b}}^*)}{\widehat{\sigma}_\varepsilon} \right), \quad S_R^* = \sum_{i=1}^n \left(\frac{r_i(\widehat{\mathbf{b}}_R^*)}{\widehat{\sigma}_\varepsilon} \right). \tag{14}$$

Finally, note that frequently the fact is used that the distribution of $(s - 1)T$ converges in law to a χ^2 distribution with $s - 1$ degrees of freedom for $n \rightarrow \infty$.

Similarly, it might be of interest if some of the (non-compositional) explanatory variables do not influence the compositional response (Type 1); in case of a

Type 3 model we could consider the problem in terms of subcompositional independence again. Possible testing of subcompositional independence in the compositional response thus should be performed directly with the involved subcomposition. For the $(D - 1) \times (P + 1)$ matrix of regression coefficients \mathbf{B}^* the null hypothesis can now be expressed as $\mathbf{A}\mathbf{B}^* = \mathbf{0}$ with the alternative that this equality does not hold. Matrix \mathbf{A} has the same structure as before, it is just adapted to the new notation, thus having $s - 1$ (internal subcompositional independence) or s (external subcompositional independence) columns, respectively, and $D - 1$ rows. The usual strategy is to employ the likelihood ratio test with the statistic (for the case of internal subcompositional independence)

$$\Lambda = \left(\frac{\det(\widehat{\boldsymbol{\Sigma}}_{\mathbf{b}})}{\det(\widehat{\boldsymbol{\Sigma}}_{\mathbf{b}_R})} \right)^{n/2},$$

used in the modified form

$$T_M = \left[n - P - 1 - \frac{1}{2}(D - (s - 1)) \right] \ln \left(\frac{\det(\widehat{\boldsymbol{\Sigma}}_{\mathbf{b}_R})}{\det(\widehat{\boldsymbol{\Sigma}}_{\mathbf{b}})} \right), \quad (15)$$

where $\widehat{\boldsymbol{\Sigma}}_{\mathbf{b}_R}$ denotes the estimated covariance matrix of the estimated matrix of regression parameters in the submodel, formed under the null hypothesis. For $n \rightarrow \infty$ the statistic T_M converges to a χ^2 distribution with $(D - 1)(s - 1)$ degrees of freedom (Johnson and Wichern 2007).

4 Robust MM Estimation

Many proposals for robust regression are available in the literature (see Maronna et al. 2006). The choice of an appropriate estimator depends on different criteria. First of all, the estimator should have desired robustness properties, i.e. robustness against a high level of contamination, and at the same time high statistical efficiency. MM estimators for regression possess the maximum breakdown point of 50% (i.e. at least 50% of contaminated samples are necessary in order to make the estimator useless), and they have a tunable efficiency. Although other regression estimators also achieve a high breakdown point, like the LTS regression estimator, their efficiency can be quite low (Maronna et al. 2006). Another criterion for the choice is the availability of an appropriate implementation in software packages. MM estimators for regression are available in the software environment R (R Development Core Team 2019). For univariate response (Type 2 regression) we refer to the function `lmrob` of the R package `robustbase` (Maechler et al. 2018), for multivariate response (Types 1 and 3) there is an implementation in the package `FRB`, which also provides inference statistics using the fast robust bootstrap (Van Aelst and Willems 2013).

4.1 MM Estimation in Regression Type 2

In the following we provide a brief description of MM estimators for regression, with the same notation as in Sect. 3.1, considering first the case of Type 2 regression. A regression M estimator is defined as

$$\hat{\mathbf{b}}^* = \operatorname{argmin}_{\mathbf{b}} \sum_{i=1}^n \rho \left(\frac{r_i(\mathbf{b})}{\hat{\sigma}_\varepsilon} \right), \tag{16}$$

where $\hat{\sigma}_\varepsilon$ is a robust scale estimator of the residuals (Maronna et al. 2006). The function $\rho(\cdot)$ should be bounded in order to achieve good robustness properties of the estimator (for details, see Maronna et al. 2006). An example is the bisquare family, with

$$\rho(r) = \begin{cases} \left(\frac{r}{k}\right)^2 \left(3 - 3\left(\frac{r}{k}\right)^2 + \left(\frac{r}{k}\right)^4\right) & \text{for } |r| \leq k \\ 1 & \text{else} \end{cases}. \tag{17}$$

The constant k is a tuning parameter which gives a tradeoff between robustness and efficiency. When k gets bigger, the resulting estimate tends to LS, thus being more efficient but less robust. A choice of $k = 0.9$ leads to a good compromise with a given efficiency.

The crucial point is to robustly estimate the residual scale which is needed for the minimization problem (Eq. 16). This can be done with an M-estimator of scale, defined as the solution of the implicit equation

$$\frac{1}{n} \sum_{i=1}^n \rho_1 \left(\frac{r_i(\mathbf{b})}{\hat{\sigma}_\varepsilon} \right) = d. \tag{18}$$

Here, ρ_1 is a bounded function and d is a constant. With this choice, regression S-estimators are defined as

$$\hat{\mathbf{b}} = \operatorname{argmin}_{\mathbf{b}} \hat{\sigma}_\varepsilon(r_1(\mathbf{b}), \dots, r_n(\mathbf{b})). \tag{19}$$

Regression S-estimators are highly robust but inefficient. However, one can compute an S-estimator $\hat{\mathbf{b}}_{(0)}^*$ as a first approach to \mathbf{b}^* , and then compute $\hat{\sigma}_\varepsilon$ as an M-estimator of scale using the residuals from $\hat{\mathbf{b}}_{(0)}^*$ (see Maronna et al. 2006). Yohai (1987) has shown that the resulting MM estimator $\hat{\mathbf{b}}^*$ inherits the breakdown point of $\hat{\mathbf{b}}_{(0)}^*$, but its efficiency under normal distribution is determined by tuning constants. The default implementation of the R function `lmrob` attains a breakdown point of 50% and an asymptotic efficiency of 95% (Maechler et al. 2018).

4.2 MM Estimation in Regression Types 1 and 3

For the case of regression Types 1 and 3, multivariate regression MM estimators can be used as robust counterparts to LS estimators. With notation as in Sect. 3.2,

compositional MM estimators are defined as

$$(\widehat{\mathbf{B}}^*, \widehat{\mathbf{C}}) = \operatorname{argmin}_{\mathbf{B}} \sum_{i=1}^n \rho \left(\frac{\mathbf{r}_i^*(\mathbf{B})^t \mathbf{C}^{-1} \mathbf{r}_i^*(\mathbf{B})}{\widehat{\sigma}} \right), \tag{20}$$

with the scale estimator $\widehat{\sigma} := \det(\widehat{\Sigma}_S)^{1/(2D-2)}$, where $\widehat{\Sigma}_S$ is obtained from a multivariate regression S-estimator (see Van Aelst and Willems 2013, for details). The estimated residual covariance matrix is then given by $\widehat{\Sigma}_\varepsilon = \widehat{\sigma}^2 \widehat{\mathbf{C}}$.

4.3 MM Testing of Subcompositional Independence

Robust hypothesis tests in linear regression are not straightforward, because they have to involve robust residuals, and some tests also rely on a robust estimation of the covariance matrix of the regression coefficients. In the following we will focus on tests which can cope with subcompositional independence.

For the univariate case (Type 2) a robust equivalent to the test mentioned in Sect. 3.3 is available. It is a likelihood ratio-type test which, unlike a Wald-type test, does not require the estimation of the covariance matrix of $\widehat{\mathbf{b}}^*$. The hypothesis to be tested is the same as stated in Sect. 3.3, namely $\mathbf{A}\mathbf{b}^* = \mathbf{0}$, with $\mathbf{A} = (\mathbf{0}, \mathbf{I})$ and \mathbf{I} an identity matrix of order $s - 1$. For the alternative hypothesis $\mathbf{A}\mathbf{b}^* \neq \mathbf{0}$. In analogy to the terms in (14), the test is based on

$$S^* = \sum_{i=1}^n \rho \left(\frac{r_i(\widehat{\mathbf{b}}^*)}{\widehat{\sigma}_\varepsilon} \right), \quad S_R^* = \sum_{i=1}^n \rho \left(\frac{r_i(\widehat{\mathbf{b}}_R^*)}{\widehat{\sigma}_\varepsilon} \right), \tag{21}$$

where $\rho(\cdot)$ is a bounded function and $\widehat{\sigma}_\varepsilon$ is a robust scale estimator of the residuals, see also Eq. (16). With the choice

$$\xi = \frac{\sum_{i=1}^n (\psi'(r_i(\widehat{\mathbf{b}}^*)/\widehat{\sigma}_\varepsilon))}{\sum_{i=1}^n \psi \left((r_i(\widehat{\mathbf{b}}_R^*)/\widehat{\sigma}_\varepsilon)^2 \right)},$$

where $\psi = \rho'$, the test statistic

$$T = \xi(S_R^* - S^*) \tag{22}$$

approximates a χ^2 distribution with $s - 1$ degrees of freedom, χ_{s-1}^2 (see Hampel et al. 1986). The null hypothesis is rejected at the significance level α if the value of the test statistic $t > \chi_{s-1}^2(1 - \alpha)$.

For regression Type 1 and 3 we can use the robust equivalent of the likelihood ratio test mentioned in Sect. 3.3. According to Eq. (15), the covariance matrix of the estimated matrix of regression parameters is needed. This can be obtained by bootstrap as follows. In their R package FRB, Van Aelst and Willems (2013) provide functionality for inference statistics in multivariate MM regression by using the idea of fast

and robust bootstrap (Salibian-Barrera and Zamar 2002). A usual bootstrap procedure would not be appropriate for robust estimators, since it could happen that a bootstrap data set contains more outliers than the original one due to an over-representation of outlying observations, thus causing breakdown of the estimator. Moreover, recalculating the robust estimates for every sample would be very time consuming. The idea of fast and robust bootstrap (FRB) is to estimate the parameters only for the original data. Let $\hat{\theta}$ contains all estimates $\hat{\mathbf{B}}$ and $\hat{\Sigma}_\varepsilon$ in vectorized form, and denote by Ω_θ the set of possible values of this vectorized model parameter. MM-estimators can be written in form of a system of fixed-point equations, i.e. thanks to a function $\mathbf{g}: \Omega_\theta \rightarrow \Omega_\theta$ such that $\hat{\theta} = \mathbf{g}(\hat{\theta})$. Indeed, if the function \mathbf{g} is known, one can estimate θ as the fixed point of the equation. The function \mathbf{g} depends on the sample, hence for a bootstrap sample we obtain a different function \mathbf{g}_b . The idea is thus to use the original estimate and the fixed-point equation for the bootstrap sample, obtaining $\hat{\theta}_b^1 := \mathbf{g}_b(\hat{\theta})$. This results in an approximation of the bootstrap estimates $\hat{\theta}_b$ which would be obtained directly from the bootstrap sample, i.e. solving $\hat{\theta}_b = \mathbf{g}_b(\hat{\theta}_b)$. Applying a Taylor expansion, an improved estimate $\hat{\theta}_b^I$ can be derived, estimating the same limiting distribution as $\hat{\theta}_b$, and being consistent for $\hat{\theta}$. For more details concerning fast and robust bootstrap for the MM-estimator of regression see Salibian-Barrera et al. (2008).

5 Case Study: the GEMAS Data Set

5.1 General Information

The GEMAS (“Geochemical Mapping of Agricultural and grazing land Soil”) soil survey geochemical campaign was conducted at European level, coordinated by EuroGeoSurveys, the association of European Geological Surveys (Reimann et al. 2014a, b). It covered 33 countries, and it focuses on those land uses that are vital for food production. The area was sampled at a density of around 1 site per 2,500 km². Samples were taken from agricultural soils (0 to 20 cm) and grazing land soils (0 to 10 cm). At each site, 5 samples at the corners and in the center of a square with 10 by 10 m were collected, and the composite sample was analyzed. Around 60 chemical elements were obtained in samples of both kinds of soil. Soil textural composition was also analyzed, i.e. the weight % of sand, silt and clay. Some parameters describing the climate (climate type, mean temperature or average annual precipitation) and the background geology (rock type) are also available. More specifically, the average annual precipitation and the annual mean temperature at the sample locations are taken from Reimann et al. (2014a) and originate from www.worldclim.org. The subdivision of the GEMAS project area into climate zones goes back to Baritz et al. (2005).

From the several variables available, we focus on the effects between the soil composition (either its chemistry or its sand-silt-clay texture) and the covariables: annual average precipitation, soil pH (both as continuous variables) and climate zones [as categorical variable, with the respective sample sizes; the categories are Mediterranean (Medi, 438), Temperate (Temp, 1,102), Boreal–Temperate (BoTe, 352) and Supraborreal (Spbo, 203)]. Figure 1 shows a set of descriptive diagrams of these variables and compositions. A total of $n = 2095$ samples of the GEMAS data set were used,

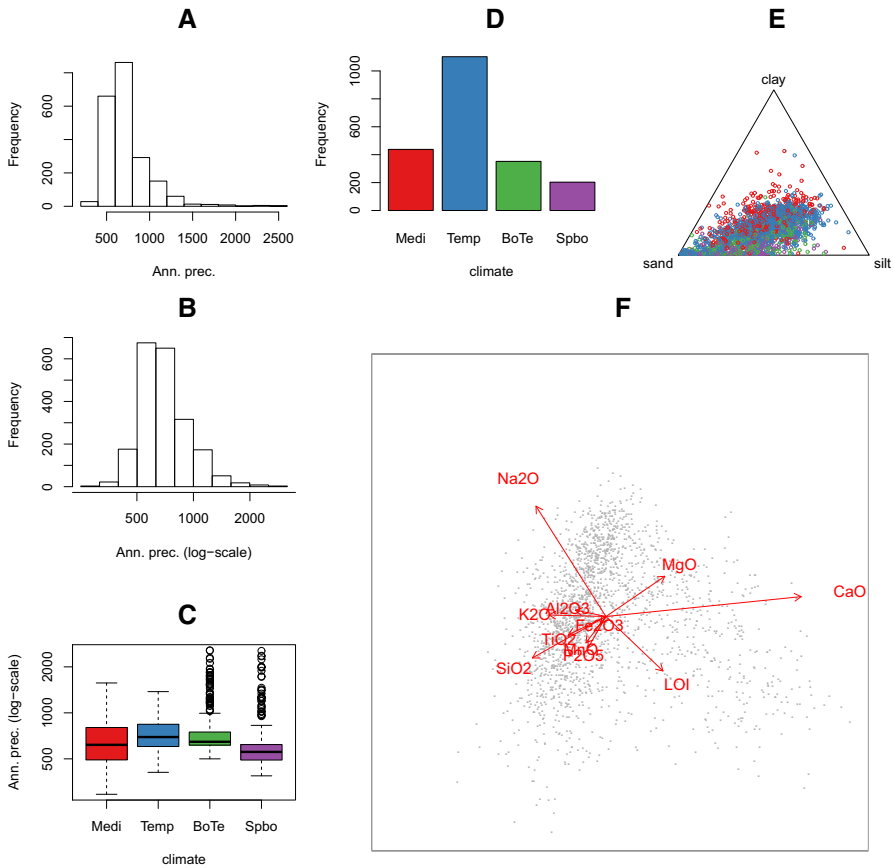


Fig. 1 Descriptive diagrams of the sets of variables used. Displayed are histograms of Annual Precipitation in the original (a) and logarithmic (b) scales. Boxplots of log-scaled Annual Precipitation (c) and histogram of sample sizes (d) according to climate groups follow. The climate groups are used to color sand-silt-clay compositions in the ternary diagram (e). Finally, the multivariate data structure of chemical compositions is captured using the compositional biplot (f)

covering almost all Europe, excepting Romania, Moldova, Belarus, Eastern Ukraine and Russia (Fig. 2). From a comparison between panels A and B (Fig. 1), one can conclude that the logarithm of Annual Precipitation is required for further treatment. Though symmetry or normality are not attained, even with a logarithm [both p values of the Anderson–Darling test for normality (Anderson and Darling 1952) were zero], at least a view by the four climatic groups suggest that departures from symmetry are moderate to mild (Fig. 1c), not going to affect negatively further regression results. As indicated above, the data present a rather unbalanced design with respect to climatic areas (Fig. 1d), particularly due to the dominance of temperate climate, which accounts for more than 50% of the samples, see also Fig. 2.

The sand-silt-clay textural composition is represented in Fig. 1e as a ternary diagram, with colors after the four climatic zones: these show a certain control on the

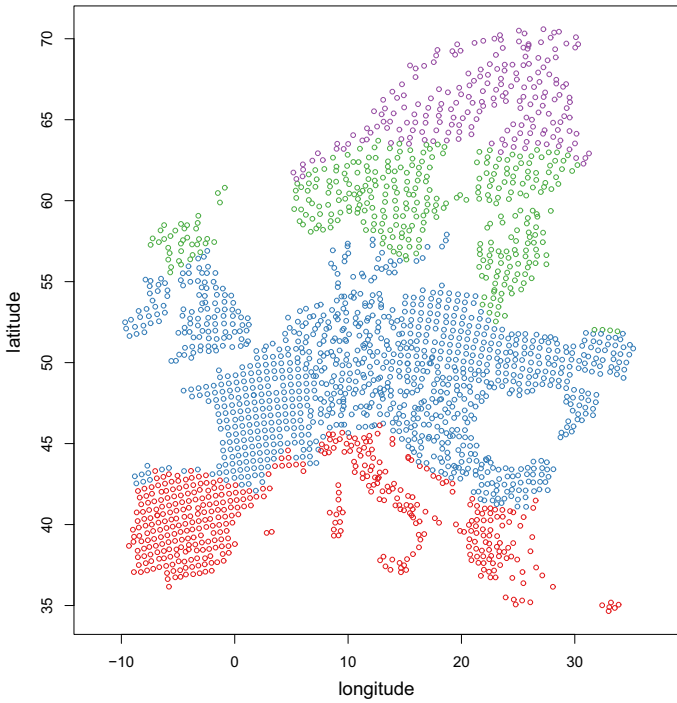


Fig. 2 Sample location. Colors after climatic zones: red for Mediterranean, blue for Temperate, green for Boreal–Temperate and violet for Supraboreal

amount of clay, and this will be explored later. With regard to the major oxide composition including SO_3 and LOI (loss on ignition), this is represented in Fig. 1f as a centered logratio covariance biplot, as conventional in compositional data analysis (Aitchison 1997; Aitchison and Greenacre 2002). This shows a quite homogeneous data set without any strong grouping that could negatively affect the quality of the next regression steps.

5.2 Grain Size Composition Versus Precipitation (Type 1 Regression)

The first task is to express the sand-silt-clay composition (response) as a function of precipitation, in log-scale (explanatory variable), using the regression model from Sect. 2.2.1. Figure 3 displays this dependence, by plotting each of the three possible logratios on the vertical axis against the logarithm of annual precipitation on the horizontal axis. The available data were transformed after the following matrix resp. these ilr coordinates

$$\mathbf{V}^t = \begin{pmatrix} \frac{-1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} & \frac{+2}{\sqrt{6}} \\ \frac{-1}{\sqrt{2}} & \frac{+1}{\sqrt{2}} & 0 \end{pmatrix}, \quad \begin{matrix} y_1^* = \frac{1}{\sqrt{6}} \ln \frac{y_{clay}^2}{y_{silt} \cdot y_{sand}} \\ y_2^* = \frac{1}{\sqrt{2}} \ln \frac{y_{silt}}{y_{sand}} \end{matrix}, \quad (23)$$

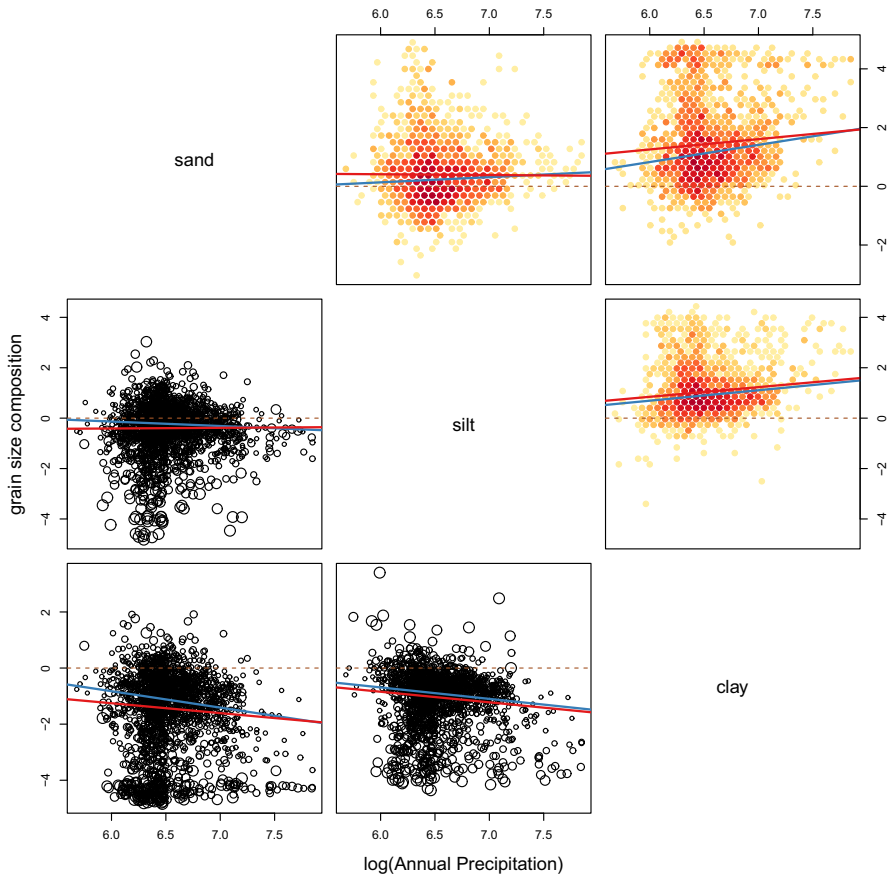


Fig. 3 Grain size composition as a function of (log) annual precipitation; observed GEMAS data (dots) and fitted models (red: classical; blue: robust). Symbol size (in the lower left panels) is inversely proportional to the weights computed in robust regression

and a model of the form of Eq. (9) was fitted by the LS method. Table 1 shows the logratio coefficients, as well as their values once back-transformed. Note that the back-transformed values would be exactly the same, whatever other logratio transformation would have been used for the calculations.

Table 1 reports the coefficients for the ilr coordinates defined in Eq. (23), the corresponding p values, and the back-transformed coefficients, using LS and MM estimators. The LS estimates show that the ratio silt-to-sand is not affected by annual precipitation, while their relation to clay does depend on this covariable. In contrast, for the MM estimators both coordinates are affected by annual precipitation. Figure 3 shows both the LS and MM models, re-expressed in each of the possible pairwise logratios. Note that the slope and intercept given for the coordinate y_2^* in Table 1 correspond to panel (2,1) of this Fig. 3. The intercepts and slopes for each of the other panels can be obtained by transforming the coefficients $(y_{\text{sand}}, y_{\text{silt}}, y_{\text{clay}})$ accordingly.

Table 1 Regression models of grain size composition against (log) annual precipitation, using LS and MM regression

	y_1^*	y_2^*	p value. y_1^*	p value. y_2^*	Sand	Silt	Clay
LS: Intercept	0.931	-0.402	0.023	0.234	0.255	0.144	0.600
LS: log(AnnPrec)	-0.298	0.019	0.000	0.718	0.366	0.376	0.258
MM: Intercept	1.813	0.656	0.000	0.014	0.055	0.139	0.806
MM: log(AnnPrec)	-0.405	-0.125	0.000	0.004	0.418	0.350	0.233

The columns refer to the estimated parameters for the ilr coordinates, the corresponding p values, and the back-transformed regression coefficients

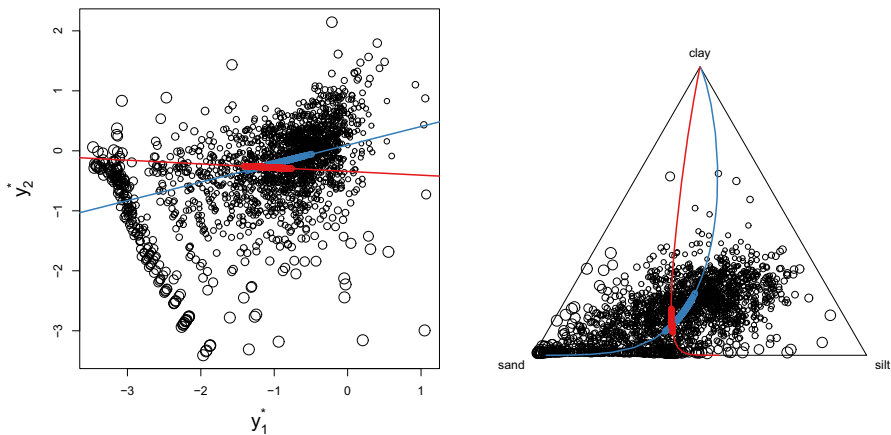


Fig. 4 Scatter diagrams (left: ilr plane; right: ternary diagram) of the data and of the predictions of both the LS (red) and MM (blue) models. The lines show the models extrapolated beyond the range of observed annual precipitation. Symbol size is inversely proportional to the weights computed by robust regression

Figure 4 shows the model predictions for the classical (red) and the robust (blue) model. The left plot presents the predictions in the ilr coordinates, as they are used in the regression models, and the right plot shows the predictions for the original composition. The symbol sizes are inversely proportional to the weights from robust MM regression, and here it gets obvious that due to very small values of clay (rounded values), data artifacts are produced in the ilr coordinates, but these observations are downweighted by MM regression. This is the main reason for the difference between the LS and the MM model.

5.3 Grain Size Composition Versus Climate (Type 1 Regression)

A regression of the grain size composition (response) against climate zones (explanatory variables) should take into account that the climate zones are ordered in a clear sequence from Mediterranean (Medi), Temperate (Temp), Boreal–Temperate (BoTe) to Supraboreal (Spbo), ordered from South to North. This is clearly seen in Fig. 5, showing a relatively constant average sand/silt ratio across climatic zones, but a clear

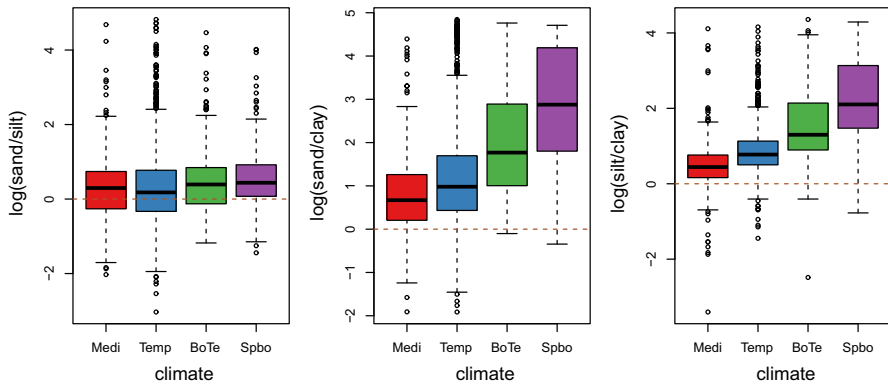


Fig. 5 Logratios of the grain size composition for the different climate zones

Table 2 Compositional centers in each climate category, expressed in percentages

	Sand	Silt	Clay
Medi	45.7117	33.2898	20.9985
Temp	49.5057	34.2782	14.0220
BoTe	58.8528	36.5830	7.6181
SpBo	65.0806	36.4010	3.7676

monotonous trend of average sand/clay and silt/clay ratios northwards. Such a trend is followed also by the compositional centers (Pawlowsky-Glahn et al. 2015) for the respective climate categories, see Table 2.

Thus the following hypothesis of uncorrelation appear as sensible:

1. the balance of sand to silt is uncorrelated with climate (i.e. the sand-silt subcomposition is internally uncorrelated with climate)
2. the balance of clay to the other two depends on climate only in so-called *linear* terms, as explained in the next paragraph.

Given these hypotheses, the same ilr coordinates as in the preceding section (Eq. 23) will be used here.

In R—package `stats`; (R Development Core Team 2019)—, a regression model with an ordered factor of 4 levels requires building an accessory ($n \times 3$)-element design or contrast matrix \mathbf{X} , where each row is taken as the corresponding row of Table 3. The labels L—“Linear”, Q—“Quadratic” and C—“Cubic” stand for the kind of trend between the four categories fitting the data, L implying that the differences between two consecutive categories are constant (Simonoff 2003).

Table 4 summarizes the numerical output of this regression model, including estimated coefficients (intercept, and effects L, Q and C) for each of the two balances, the p values of the hypotheses of null coefficient, and the back-transformed coefficients. These results are given for both classical (LS) and robust (MM) regression. Classical LS regression shows that C and Q effects can be discarded for y_2^* but not L effects, i.e. the first hypothesis (inner uncorrelation of the sand-silt subcomposition with climate) must be rejected. With regard to the second hypothesis, nullifying the

Table 3 Row vectors to construct the design matrix associated with the categorical variable *climate*

	L	Q	C
Medi	-0.6708	0.5000	-0.2236
Temp	-0.2236	-0.5000	0.6708
BoTe	0.2236	-0.5000	-0.6708
SpBo	0.6708	0.5000	0.2236

These numbers result from applying the R function `contrasts` on the ordered factor variable *climate*. L stands for linear effect, Q for quadratic effect and C for cubic effect

Table 4 Fitted coefficients and *p* values of the regression models of grain size composition versus climate

	y_1^*	y_2^*	<i>p</i> value. y_1^*	<i>p</i> value. y_2^*	Sand	Silt	Clay
LS: intercept	-1.238	-0.308	0.000	0.000	0.548	0.355	0.097
LS: L	-1.195	-0.142	0.000	0.001	0.493	0.403	0.103
LS: Q	-0.120	-0.019	0.002	0.594	0.354	0.344	0.302
LS: C	0.045	0.009	0.165	0.765	0.325	0.329	0.346
MM: intercept	-1.143	-0.223	0.000	0.000	0.515	0.376	0.109
MM: L	-1.202	-0.157	0.000	0.000	0.498	0.399	0.102
MM: Q	-0.270	-0.085	0.000	0.006	0.390	0.346	0.264
MM: C	0.017	0.054	0.728	0.022	0.319	0.344	0.338

coefficients for L and Q effects on y_1^* are significantly different from zero (*p* values smaller than 0.05 critical level), which implies that the second hypothesis is false as well. Nevertheless, C effects can be discarded. A global test in the fashion of what was explained in Sect. 3.3 gives a zero *p* value for the hypothesis of absence of Q or C effects, thus supporting these conclusions. Robust regression delivers a similar picture, except that here all effects are significant for y_2 .

Of course, other contrasts could be used for this analysis, depending on the nature of the hypotheses of dependence that we are interested in testing. If, for instance, one would want to check whether soils from different climatic zones have on average the same soil texture, one could have used the `constr.treatment` function of R to force this sort of comparison.

One way or another, in a categorical regression model like this, the intercept can be interpreted as a sort of global average value compensating for the lack of balance between the four categories. While the conventional compositional center is [sand; silt; clay] = [52.39%; 35.27%; 12.34%] the least squares regression delivers an estimate [54.83%; 35.48%; 9.69%] and the robust regression [51.54%; 37.60%; 10.86%], both downweighting the importance of clay. Note that this intercept does not depend on which contrast set is chosen for capturing the categorical variable.

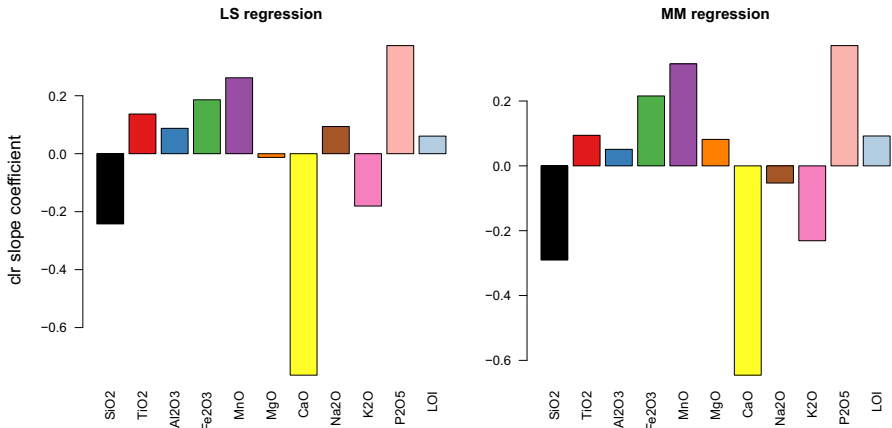


Fig. 6 Least squares (left) and MM (right) regression coefficients of the clr transformed major oxide composition versus log-annual precipitation

5.4 Major Oxides Versus Precipitation (Type 1 Regression)

Much richer hypotheses can be tested if the composition used has many parts. To illustrate this, a regression of the major oxides against (log) annual precipitation follows. A natural initial question is whether the soil geochemistry is influenced by precipitation. For this purpose, Fig. 6 shows the clr coefficients estimated with classical least squares and robust regression: clr coefficients should never be interpreted individually; rather, differences between them can be understood as the influence on a particular pairwise logratio. Thus, we are seeking for the smallest differences between coefficients as they identify pairs of variables whose balance is not influenced by the explanatory variable. As a subset of pairwise logratios identifies a (sub)composition, this gives information about subcompositions that might be potentially internally independent of the covariable, such as:

- $\text{TiO}_2\text{--Fe}_2\text{O}_3\text{--MnO}$
- $\text{Al}_2\text{O}_3\text{--LOI}$ (with Na_2O in least squares regression, or MgO in robust regression)
- $\text{SiO}_2\text{--K}_2\text{O}$

A set of ilr coordinates is selected accordingly to contain balances between these subcompositions. The matrix of signs to build these balances is given in Table 5. Remember that in a sign table, +1 indicates variables that appear in the numerator of the balance, -1 variables in the denominator, and 0 variables are not involved in that particular balance. For instance, the balance between the subcompositions $\text{TiO}_2\text{--Fe}_2\text{O}_3\text{--MnO}$ and $\text{Al}_2\text{O}_3\text{--Na}_2\text{O--LOI}$ is y_4^* , and the balances (y_7^*, y_8^*) describe the internal variability in the subcomposition $\text{TiO}_2\text{--Fe}_2\text{O}_3\text{--MnO}$.

Using this set of balances, a regression model with explanatory variable (log) annual precipitation is fit, with LS and MM regression. Results are reported in Table 6. Paying attention to the p values of the slopes of the two models, we conclude that the subcomposition $\text{Al}_2\text{O}_3\text{--Na}_2\text{O--LOI}$ (y_7^*, y_8^*) is internally independent of annual precipitation (both classical and robust methods agree in that). Loosely speaking, the

Table 5 Table of signs to construct the balances for major oxide composition in a regression problem against (log) annual precipitation

	y_1^*	y_2^*	y_3^*	y_4^*	y_5^*	y_6^*	y_7^*	y_8^*	y_9^*	y_{10}^*
SiO ₂	+1	-1	0	0	0	0	0	0	+1	+1
TiO ₂	+1	+1	+1	+1	+1	+1	0	0	0	0
Al ₂ O ₃	+1	+1	+1	-1	0	0	+1	+1	0	0
Fe ₂ O ₃	+1	+1	+1	+1	+1	-1	0	0	0	0
MnO	+1	+1	+1	+1	-1	0	0	0	0	0
MgO	-1	0	0	0	0	0	0	0	0	0
CaO	+1	-1	0	0	0	0	0	0	-1	0
Na ₂ O	+1	+1	+1	-1	0	0	+1	-1	0	0
K ₂ O	+1	-1	0	0	0	0	0	0	+1	-1
P ₂ O ₅	+1	+1	-1	0	0	0	0	0	0	0
LOI	+1	+1	+1	-1	0	0	-1	0	0	0

Table 6 Intercept (int) and slope (slp) estimated coefficients and p values (p), for least squares (LS) and robust (MM) regression

	LS.int	MM.int	LS.slp	MM.slp	LS.int.p	MM.int.p	LS.slp.p	MM.slp.p
y_1^*	0.596	1.229	0.013	-0.086	0.053	0.000	0.776	0.053
y_2^*	-7.722	-7.469	0.822	0.789	0.000	0.000	0.000	0.000
y_3^*	3.301	3.454	-0.218	-0.233	0.000	0.000	0.000	0.000
y_4^*	-3.450	-3.937	0.140	0.218	0.000	0.000	0.000	0.000
y_5^*	2.876	3.192	-0.082	-0.130	0.000	0.000	0.003	0.000
y_6^*	-1.011	-0.654	-0.035	-0.086	0.000	0.000	0.085	0.000
y_7^*	-1.151	-0.339	0.024	-0.076	0.002	0.361	0.662	0.157
y_8^*	1.752	1.176	-0.004	0.074	0.000	0.000	0.911	0.089
y_9^*	-1.243	-0.242	0.451	0.314	0.036	0.000	0.761	0.000
y_{10}^*	2.825	2.791	-0.044	-0.042	0.000	0.000	0.111	0.073

same applies to the balances SiO₂/K₂O (y_{10}^*) and MgO against all other components (y_1^*). Finally the balance TiO₂/Fe₂O₃ (y_6^*) appears to be uncorrelated with annual precipitation only from a least-squares perspective.

Now, global tests of internal and external independence of Al₂O₃-Na₂O-LOI with respect to annual precipitation were performed after the methodology of Sect. 3.3, and delivered p values of 0.884 and 0, respectively. These results are somehow at odds with the common understanding of weathering as a process of enrichment in Al₂O₃ (and perhaps LOI) at the expenses of Na₂O (and CaO). Annual precipitation, one of the factors of chemical weathering, is not showing any significant effect on the logratio Al₂O₃/Na₂O. The robust global test, on the other hand, results in significant effects: in both cases, the p values are zero.

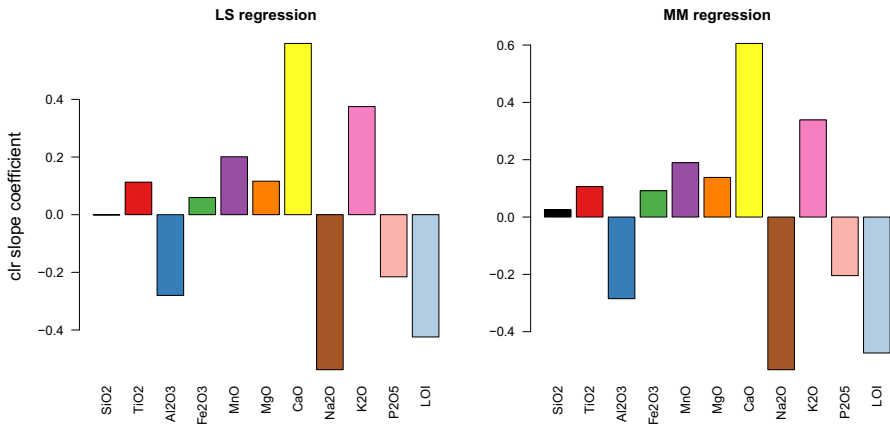


Fig. 7 Regression gradient of pH against the major oxide composition, expressed in clr coefficients: least squares estimates (LS, left) and robust estimates (MM, right)

5.5 Compositional Input: pH Versus Major Oxides (Type 2 Regression)

A similar approach can be followed if the goal is to predict an external quantitative variable as a function of a composition. This section presents the regression of soil pH against the major oxide composition, the rationale being that the mineral composition of a soil may be influenced—for example through accelerating dissolution of unstable minerals under certain pH values at atmospheric conditions—resp. that the mineralogy may be one of the factors controlling availability of free H_3O^+ ions—for example soils on karstic landscapes are strongly buffered, while those on top of felsic rocks are strongly acidified. Following the same steps as in the preceding subsection, it is convenient to start having a look at the regression model coefficients expressed in clr. In that case, though, it is necessary to start the analysis with a black-box ilr coordinate set, and convert the estimated ilr coefficients to clr coefficients, in order to avoid the singularity of the clr variance matrix. Similar as in Fig. 6, one should pay attention to bars with similar length, in order to identify pairs of variables which balance has no influence on the explained variable. Figure 7 suggests the following pairs, which are taken into account by constructing balances according to Table 7: $\text{Al}_2\text{O}_3/\text{P}_2\text{O}_5$ (y_6^*) and $\text{Ti}_2\text{O}_5/\text{MgO}$ (y_{10}^*). Moreover, by taking the variables with the largest positive and negative weights we also find balances which could concentrate most of the predicting power for pH: these are $\text{CaO}/\text{Na}_2\text{O}$ (y_2^*) and $\text{K}_2\text{O}/\text{LOI}$ (y_5^*).

The regression results are presented in Table 8 for LS and MM regression. The table shows the estimated regression coefficients and the corresponding p values. Both methods reveal that the coefficients for balances y_2^* and y_5^* are significant, while those for balances y_6^* and y_{10}^* are not. Using the methods from Sect. 3.3 we can further test subcompositional independence of pH with respect to subcompositions such as $\text{MnO}-\text{Fe}_2\text{O}_3-\text{MgO}-\text{TiO}_2$ (p value = 0.291, subcompositionally independent), or $\text{P}_2\text{O}_5-\text{Al}_2\text{O}_3-\text{K}_2\text{O}-\text{LOI}$ (p value = $6.75\text{e}-43$, not independent).

As an example, Fig. 8 investigates more deeply the relationship between pH and balance y_2^* , which relates CaO and Na₂O, where the color information is for high (red)

Table 7 Table of signs to construct the balances for major oxide composition in a regression problem with pH as the response variable

	y_1^*	y_2^*	y_3^*	y_4^*	y_5^*	y_6^*	y_7^*	y_8^*	y_9^*	y_{10}^*
SiO ₂	-1	0	-1	0	0	0	-1	0	0	0
TiO ₂	-1	0	-1	0	0	0	1	1	1	1
Al ₂ O ₃	-1	0	1	-1	0	1	0	0	0	0
Fe ₂ O ₃	-1	0	-1	0	0	0	1	1	-1	0
MnO	-1	0	-1	0	0	0	1	-1	0	0
MgO	-1	0	-1	0	0	0	1	1	1	-1
CaO	1	1	0	0	0	0	0	0	0	0
Na ₂ O	1	-1	0	0	0	0	0	0	0	0
K ₂ O	-1	0	1	1	1	0	0	0	0	0
P ₂ O ₅	-1	0	1	-1	0	-1	0	0	0	0
LOI	-1	0	1	1	-1	0	0	0	0	0

Table 8 LS and MM regression coefficients and *p* values for a regression of pH on the major oxide balances

	LS.coeff	MM.coeff	LS. <i>p</i> .value	MM. <i>p</i> .value
Intercept	7.099	7.092	0.000	0.000
y_1^*	0.044	0.057	0.077	0.037
y_2^*	0.800	0.805	0.000	0.000
y_3^*	-0.349	-0.396	0.000	0.000
y_4^*	0.223	0.178	0.020	0.077
y_5^*	0.565	0.575	0.000	0.000
y_6^*	-0.046	-0.057	0.549	0.458
y_7^*	0.111	0.095	0.011	0.038
y_8^*	-0.091	-0.068	0.058	0.165
y_9^*	0.045	0.025	0.608	0.778
y_{10}^*	-0.002	-0.022	0.966	0.685

and low (blue) values of pH. One can see that the ratio of CaO and Na₂O increases strongly with higher values of pH, leading to a non-linear relationship. The reason is seen in the middle panel of Fig. 8 using the same coloring, where high pH values are connected to high concentrations of CaO. These pH rich samples are indicated in the project area in the right panel with red color. This supports the starting hypothesis of a strong control on pH of the buffering ability of carbonate soils. This trend can be explained as the contrast between silicic-clastic plus crystalline rocks with significant contributions of Na-rich silicates versus carbonate karstic landscapes, dominated by CaCO₃ with its very strong buffering effect at slightly basic pH values. Such a complex trend could be better captured either with a non-linear regression method, or stepwise linear regression to be carried out only for the samples which behave similarly (blue or red).

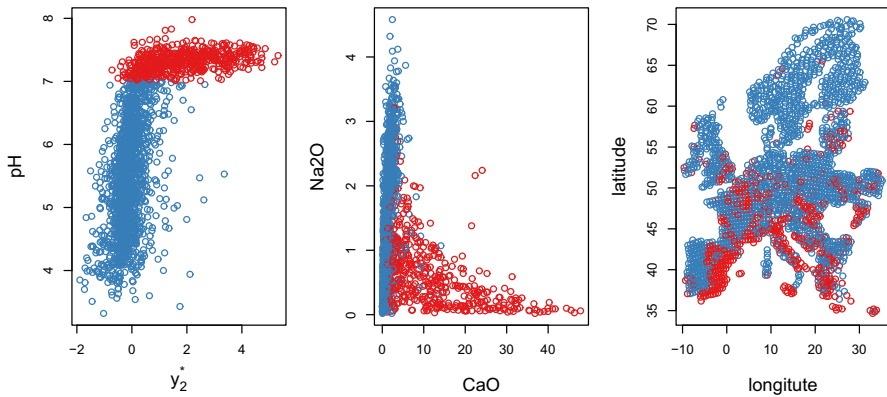


Fig. 8 pH versus balance y_2^* (left), Na_2O versus CaO (middle), and map of the area (right), with color indicating pH values lower or higher than 7

Nevertheless, it is also interesting to inspect regression diagnostic plots. For this purpose, a robust regression diagnostic plot based on standardized regression residuals and robust Mahalanobis distances was employed to unmask outlier observations. This plot enables distinguishing between vertical outliers (outliers in y -direction) and leverage points (Rousseeuw and van Zomeren 1990). The latter are observations with unusual values in the explanatory variables, which can either strengthen (good leverage points) or break (bad leverage points) the overall regression trend. The cut-off lines for vertical outliers are represented by the quantiles of the standard normal distribution and the red curve represents a kernel fit of the points. The MM regression diagnostic plot (Fig. 9, left) reveals only few vertical outliers and bad leverage points which indicates that model assumptions such as homoscedasticity (see Fig. 10) are fulfilled. However, in line with previous findings, observations with high pH values (higher than 7) form a specific pattern. This is further confirmed when observed versus fitted values are plotted (Fig. 9, right) which indicates a certain bias resulting from an underlying non-linearity in the relationships between the response and the explanatory variables, and this was already assumed previously.

5.6 Grain Size Versus Chemistry (Type 3 Regression)

In a final example we want to investigate if the grain size composition is affected by the major oxides. Similar as in the previous example, the modelling hypothesis here is that the grain size is controlled by mineralogy, with certain minerals (and their constituting elements) being enriched in certain fractions: coarse quartz (SiO_2) and feldspars ($(\text{K},\text{Na},\text{Ca})\text{Al}(\text{Al},\text{Si})_2\text{O}_8$) in sand; highly stable heavy minerals as rutile (TiO_2) or apatite ($\text{Ca}_5(\text{PO}_4)_3(\text{F},\text{Cl},\text{OH})$) in silts; and alteration and clay minerals with highly complex chemistry (but typically enriched in K, Fe, Mg, Al, OH and water) in clay-sized fractions. These relations, combined with the fact that size information is of relatively bad analytical quality (see Table 9) one could consider the possibility to support the quantification of size particles with chemistry.

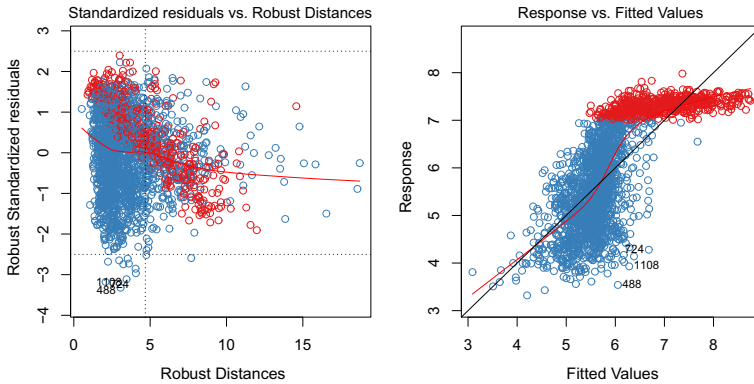


Fig. 9 Robust regression diagnostic plot (left), observed versus fitted pH values using the major oxides as predictors (right). Observations with pH values higher than 7 are marked in red

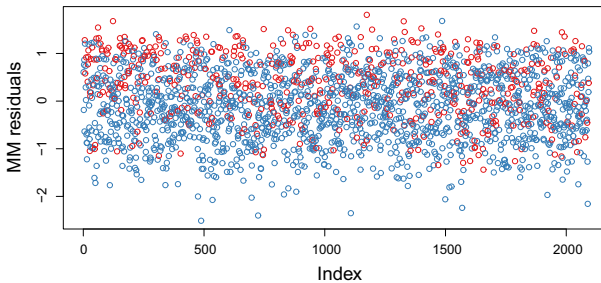


Fig. 10 MM regression residuals. Observations with pH values higher than 7 are marked in red

Table 9 Descriptive statistics of the variability of the total sum of %sand + %silt + %clay, which should be 100%

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
59.00	95.00	99.00	98.97	103.00	141.00

To start with, we follow the same approach as in the previous cases and plot the coefficients resulting from LS or robust regression in terms of clr coefficients, Fig. 7. We can then look at similar contributions of the several oxides on each grain size fraction to formulate hypotheses. Note that this leads to a matrix of regression coefficients, linking the grain size distribution as responses to the major oxide composition as explanatory variables. Table 10 reports the coefficients of an LS Type 3 regression model, albeit with all coefficients expressed in clr representation.

For establishing subcompositional independence, though, it is more convenient to work in isometric logratios. Hence, and given the results obtained until now it appears sensible to study the single component independence of clay (vs. sand and silt in balance y_2^*) on the one hand, and on the other the internal subcompositional

Table 10 Least-square coefficients of a Type 3 regression model, represented in clr coefficients (first 3 columns) and in ilr coordinates (last 2 columns)

	Sand	Silt	Clay	y_1^*	y_2^*
Intercept	-3.5756	3.3797	0.1958	4.9181	0.2398
SiO ₂	0.7096	-0.4581	-0.2515	-0.8257	-0.3080
TiO ₂	-0.8453	0.5952	0.2501	1.0186	0.3063
Al ₂ O ₃	0.4882	-0.5858	0.0977	-0.7594	0.1196
Fe ₂ O ₃	-0.0012	0.0434	-0.0422	0.0315	-0.0516
MnO	-0.1958	0.0029	0.1928	0.1405	0.2362
MgO	-0.0402	-0.1334	0.1737	-0.0659	0.2127
CaO	-0.0794	-0.0214	0.1008	0.0410	0.1235
Na ₂ O	0.1614	0.2218	-0.3832	0.0428	-0.4693
K ₂ O	-0.5873	0.1671	0.4203	0.5335	0.5147
P ₂ O ₅	0.3422	0.0191	-0.3613	-0.2284	-0.4425
LOI	0.0480	0.1493	-0.1973	0.0716	-0.2417

independence of the subcomposition {sand, silt} in balance y_1^* . The corresponding model coefficients are also reported in Tables 10 and 12 respectively for y_1^* and y_2^* .

To simplify the model, we now seek major oxide subcompositions that are non-influential on each of these two logratios of grain size composition. A subcomposition will not be influential on a certain response Y_k^* if the gradient (isometric) logratio coefficients \mathbf{b}_k^* associated to that subcomposition are zero, which is equivalent to saying that the clr coefficients of the gradient in that subcomposition show the same value (not necessarily zero!).

For each of the grain size ilr coefficients the heuristic to basis selection of Sect. 2.3 was applied (using the centroid clustering method as hierarchical agglomeration criterion), and an ilr basis was obtained. The balance of clay to sand-silt showed no subcompositional independence worth exploring. Better results were obtained for the balance of silt to sand. Table 11 contains the sign table defining the ilr balances, and the associated gradient coefficients with their corresponding p value of the hypothesis of null coefficient. At the global 5% level of confidence and applying an incremental Bonferroni correction, we accept the null hypothesis for all balances between x_1^* and x_5^* (this last one because $0.05/5 < 0.0110$), hence the following subcompositions have no internal influence on the sand-to-silt balance: SiO₂-Al₂O₃ and CaO-Fe₂O₃-MgO-Na₂O-LOI. The hypotheses of external independence of both subcompositions produce p values below 2×10^{-16} resp. 5×10^{-6} and hence external independence must be disregarded. These results suggest that minerals carrying most of Al₂O₃ and SiO₂ (essentially, feldspars vs. quartz) may not preferentially concentrate between sand and silt, and the same happens with dominant minerals composed of elements of the subcomposition CaO-Fe₂O₃-MgO-Na₂O-LOI (plagioclase, calcite, dolomite, silicates of smaller typical crystal size). On the contrary, elements such as TiO₂-P₂O₅-K₂O (captured by balances x_7^* to x_9^*) and x_{10}^* , the balance of SiO₂-Al₂O₃ versus the other, do have a strong influence on silt/sand. This supports the previous understanding that stable heavy minerals such as rutile or apatite should be (relatively) enriched on silt sized soils, while sandy soils should have more quartz and feldspars. Note that the

Table 11 Least squares regression model coefficients and null p values (last three columns) for a tailored lr -representation of the major oxide composition (represented in the first ten columns) to explain the logratio sand to silt (note: intercept not included)

	SiO ₂	TiO ₂	Al ₂ O ₃	Fe ₂ O ₃	MnO	MgO	CaO	Na ₂ O	K ₂ O	P ₂ O ₅	LOI	Estim.	Stderr.	p value
x_1^*	0	0	0	0	0	0	-1	1	0	0	0	0.001	0.017	0.943
x_2^*	0	0	0	-1	0	0	1	1	0	0	0	0.009	0.046	0.855
x_3^*	0	0	0	1	0	0	1	1	0	0	-1	-0.029	0.033	0.376
x_4^*	-1	0	1	0	0	0	0	0	0	0	0	0.047	0.055	0.398
x_5^*	0	0	0	1	-1	0	1	1	0	0	1	-0.084	0.033	0.011
x_6^*	0	0	0	1	1	-1	1	1	0	0	1	0.120	0.035	0.001
x_7^*	0	0	0	1	1	1	1	1	0	-1	1	0.252	0.034	0.000
x_8^*	0	-1	0	0	0	0	0	0	1	0	0	-0.343	0.043	0.000
x_9^*	0	1	0	-1	-1	-1	-1	-1	1	-1	-1	0.962	0.044	0.000
x_{10}^*	-1	1	-1	1	1	1	1	1	1	1	1	1.239	0.068	0.000

Table 12 Least squares regression model coefficients and null p values (last three columns) for a tailored lr -representation of the major oxide composition (represented in the first ten columns) to explain the balance of clay against sand and silt (note: intercept not included)

	SiO ₂	TiO ₂	Al ₂ O ₃	Fe ₂ O ₃	MnO	MgO	CaO	Na ₂ O	K ₂ O	P ₂ O ₅	LOI	Estimate	Std.error	p value
x_1^*	0	0	-1	0	0	0	1	0	0	0	0	0.566	0.055	0.000
x_2^*	0	0	0	0	-1	1	0	0	0	0	0	-0.146	0.028	0.000
x_3^*	0	0	0	0	0	0	0	-1	0	1	0	-0.192	0.030	0.000
x_4^*	-1	0	0	0	0	0	0	0	0	0	1	0.635	0.033	0.000
x_5^*	0	-1	0	0	1	1	0	0	0	0	0	-0.801	0.042	0.000
x_6^*	0	1	-1	0	1	1	-1	0	0	0	0	0.793	0.052	0.000
x_7^*	1	0	0	0	0	0	0	-1	0	-1	1	-0.284	0.028	0.000
x_8^*	0	1	1	-1	1	1	1	0	0	0	0	0.040	0.068	0.560
x_9^*	0	1	1	1	1	1	1	0	-1	0	0	-0.431	0.048	0.000
x_{10}^*	-1	1	1	1	1	1	1	-1	1	-1	-1	0.589	0.023	0.000

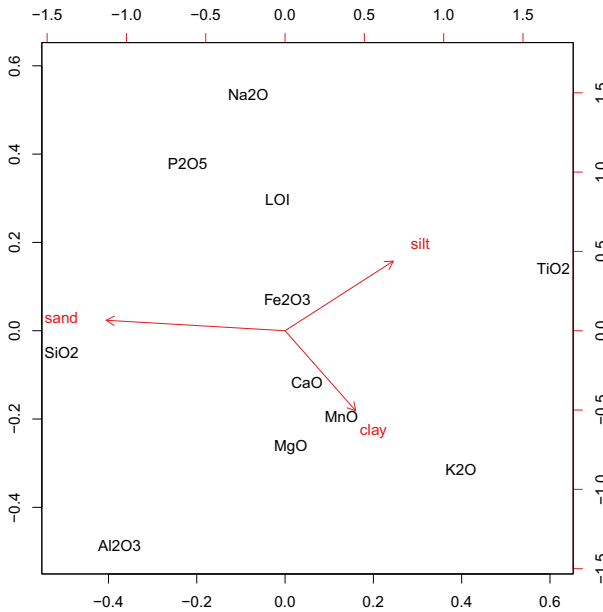


Fig. 11 Simultaneous plot of the left and (scaled) right singular vectors of the regression coefficients matrix, expressed in clr coefficients on both image and origin compositional spaces

picture for y_2^* is completely different: here we can only hope to simplify one coefficient, namely x_8^* , giving the balance of Fe with respect to the other mafic components. This is nevertheless irrelevant for the sake of subcompositional independence testing, because the rest of the balances between mafic components (x_2^* , x_5^* and x_7^*) do show significant coefficients.

Another way of looking at the model coefficients is to express them via the singular value decomposition of Eq. (11). A naive simultaneous plot of the left and right singular vectors (these last ones scaled by the singular values) is given in Fig. 11. In this diagram, links joining two variables represent the direction (on the origin or on the image simplexes, resp. \mathcal{S}_x or \mathcal{S}_y) associated with fastest change of the logratio of the two variables involved. A pair of parallel links, one involving components of \mathcal{S}_x and the other linking components of \mathcal{S}_y , suggests that the logratio between the involved response variables is strongly controlled by the logratio of the explanatory variables. For instance, the silt–clay link is reasonably parallel to the link $\text{Na}_2\text{O}–\text{Al}_2\text{O}_3$; the same can be said of the links silt–sand versus $\text{TiO}_2–\text{SiO}_2$, or of clay–sand versus $\text{K}_2\text{O}–\text{SiO}_2$. An analogous reasoning applies for orthogonal links: they indicate lack of dependence between the two sets of variables involved. In other words, by finding orthogonal links we identify subcompositions to test for potential subcompositional independence. For instance, the link sand–silt is roughly orthogonal to the sets $\text{SiO}_2–\text{Al}_2\text{O}_3$ and $\text{CaO}–\text{Fe}_2\text{O}_3–\text{MgO}–\text{Na}_2\text{O}–\text{LOI}(-\text{MnO})$, that is to the subcompositions that were previously tested. Similarly, the diagram suggests as well tests for subcompositional independence of sand–clay with respect to the subcompositions $\text{SiO}_2–\text{P}_2\text{O}_5–\text{Na}_2\text{O}$ or $\text{Al}_2\text{O}_3–\text{Fe}_2\text{O}_3(-\text{LOI})$, or even $\text{K}_2\text{O}–\text{TiO}_2$.

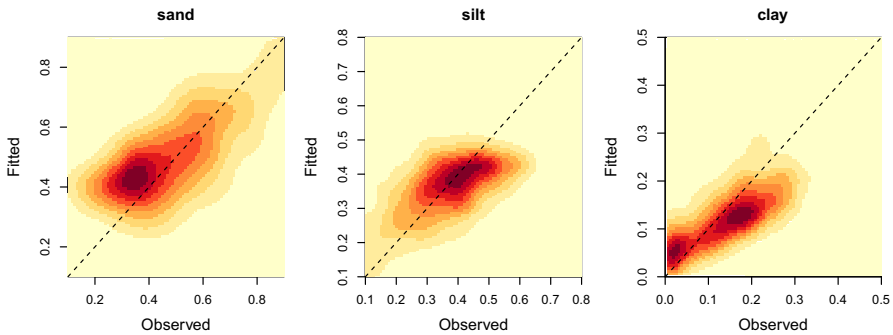


Fig. 12 Observed versus fitted values of the grain size composition, using the major oxides as predictors. Points on the dashed lines would indicate a perfect match between the observed and fitted values. The contour lines indicate the density of the point distribution

Finally, we illustrate the model in terms of the fitted values. For this purpose one can use any ilr coordinates representing the grain size composition, and any ilr coordinates representing the major oxides, and perform LS regression to obtain the fitted values in ilr coordinates. The appropriate inverse ilr transformation leads to the back-transformed fitted values of the grain size distribution, which can be compared to the observed values in Fig. 12 (as kernel density estimates). The same results will be obtained with any other logratio transformation. The linear model has at least some predictive power, and one can see a clearer relationship with sand and clay, and a weaker with silt. This suggests that the major oxides are affecting the grain size composition mainly by its sand and clay proportions. Obviously, several factors do contribute to this discrepancy, among other the information effect (the regression line of true values as a function of predicted values cannot lie above the 1:1 bisector), the presence of outliers, the bad quality of the input grain size data, the non-linearity of the back-transformation of predictions from logratios to original components, or the highly complex relations between chemistry, mineralogy and texture that form the basis to attempt such a prediction. With respect to outliers, the predictive power can be improved by using a robust estimator. The non-linearity of the back-transformation is something that can easily be corrected by means of Hermitian integration of the conditional distribution of the soil grain size composition provided by Eq. (10), as proposed by Aitchison (1986). But much more important than those effects are the uncertainty on the textural data and the complexity of the relation we are trying to capture here. Indeed, if the goal of the study would be that prediction, linear regression might not be the most appropriate technique. Tackling this complexity is a matter of predictive models, beyond the scope of this contribution.

6 Conclusions

The purpose of this contribution was to outline the concept of regression analysis for compositional data, and to show how the analysis can be carried out in practice with real data. We distinguished three types of regression models: Type 1, where

the response is a composition and the explanatory variable(s) is a (are) real non-compositional variable(s); Type 2 with a composition as explanatory variables and a real response, and Type 3 where both the responses and the explanatory variables are compositions. Note that one could also consider the case where regression is done within one composition, by splitting the compositional parts into a group forming the responses, and a group representing the explanatory variables. This case has not been treated here because it requires a so-called errors-in-variables model, see Hružová et al. (2016) for details.

For all three types of models it is essential how the composition is treated for regression modeling. A geometrically sound approach is in terms of orthonormal coordinates, so-called balances, which can be constructed in order to obtain an interpretation of the regression coefficients and for testing different hypotheses. If the interest is not in the statistical inference but only in the model fit and in the fitted values, any log-ratio coordinates would be appropriate to represent the composition. Note that the clr transformation would not be appropriate for Type 2 or Type 3 regression models, since the resulting data matrix is singular, leading to problems for the parameter estimation when the composition plays the role of the explanatory variables.

Classical least-squares (LS) regression as well as robust MM regression have been considered to estimate the regression parameters and the corresponding p values for the hypothesis tests. If the model requirements are fulfilled, the LS regression estimator is the so-called best linear unbiased estimator (BLUE) with the corresponding optimality properties (see, e.g., Johnson and Wichern 2007), but in that case also MM regression leads to an estimator with high statistical efficiency. However, in case of model violations, e.g., due to data outliers, these optimality properties are no longer valid. Still, the MM estimator is reliable because it is highly robust against outliers, both in the explanatory and in the response variables. In practical applications it might not always be clear if outliers are present in the data at hand. In this case it could be recommended to carry out both types of analysis and compare the results. In particular, one could inspect diagnostics plots from robust regression (as it was done in Sect. 5.5) in order to identify potential outliers that could have affected the LS estimator, see Maronna et al. (2006).

The different regression types and estimators have been applied to an example data set from the GEMAS project (Reimann et al. 2014a, b). All presented examples are only for illustrative purposes, but they show how balances can be constructed and how hypotheses can be tested. For the robust estimators, functions are available in the R packages *robustbase* (Maechler et al. 2018) and *FRB* (Van Aelst and Willems 2013). It is important to note that not only the regression parameters are estimated robustly with these packages, but robust estimation is also carried out for estimating the standard errors and for hypothesis testing, for the residual variance, the multiple R^2 measure, etc. We demonstrate the possibilities of regression diagnostics in Sect. 5.5. In most examples, a comparison of LS and MM regression has been provided.

An important issue in the regression context is the problem of variable selection, or subcompositional independence. In particular for Type 2 and 3 where the explanatory variables are originating from a composition, it is not straightforward how to end up with the “best subset” of compositional parts that does not contain non-informative parts and still yields a model with similar predictive power as the full model. There are

approaches available in the literature to reduce the number of components, see, e.g., Pawlowsky-Glahn et al. (2011), Hron et al. (2013), Mert et al. (2015) and Greenacre (2019). However, there are no methods of subcompositional independence which work equivalently to non-compositional methods, such as forward or backward variable selection; only a brief outlook for those in the compositional context was sketched in Filzmoser et al. (2018). Those methods will be treated in our future research.

Acknowledgements Karel Hron and Peter Filzmoser gratefully acknowledge the support by Czech Science Foundation GA19-01768S.

Funding Open access funding provided by ZHAW Zurich University of Applied Sciences

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Aitchison J (1986) The statistical analysis of compositional data. Monographs on Statistics and Applied Probability, London (UK): Chapman & Hall, London. (Reprinted in 2003 with additional material by The Blackburn Press), ISBN 0-412-28060-4
- 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, volume I, II and addendum, Barcelona (E): International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E), ISBN 84-87867-97-9, pp 3–35
- Aitchison J, Greenacre M (2002) Biplots for compositional data. *J R Stat Soc Ser C (Appl Stat)* 51(4):375–392
- Aitchison J, Barceló-Vidal C, Egozcue JJ, Pawlowsky-Glahn V (2002) A concise guide for the algebraic-geometric structure of the simplex, the sample space for compositional data analysis. In: Bayer U, Burger H, Skala W (eds) Proceedings of IAMG'02—The eighth annual conference of the International Association for Mathematical Geology, volume I and II, Selbstverlag der Alfred-Wegener-Stiftung, Berlin, pp 387–392, ISSN 0946-8978
- Anderson TW, Darling DA (1952) Asymptotic theory of certain “goodness-of-fit” criteria based on stochastic processes. *Ann Math Stat* 23:193–212
- Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V (2001) Mathematical foundations of compositional data analysis. In Ross G (ed) Proceedings of IAMG'01—The VII annual conference of the international association for mathematical geology, Cancun (Mex)
- Baritz R, Fuchs M, Hartwich R, Krug D, Richter S (2005) Soil regions of the European Union and adjacent countries 1:5,000,000 (Version 2.0)—Europaweite thematische Karten und Datensätze. European Soil Bureau Network
- Billheimer D, Guttorp P, Fagan W (2001) Statistical interpretation of species composition. *J Am Stat Assoc* 96(456):1205–1214
- Coenders G, Martín-Fernández J, Ferrer-Rosell B (2017) When relative and absolute information matter: compositional predictor with a total in generalized linear models. *Stat Model* 17(6):494–512
- Daunis-i-Estadella J, Egozcue JJ, Pawlowsky-Glahn V (2002) Least squares regression in the Simplex. In: Bayer U, Burger H, Skala W (eds) Proceedings of IAMG'02—the eighth annual conference of

- the International Association for Mathematical Geology, volume I and II, Selbstverlag der Alfred-Wegener-Stiftung, Berlin, ISSN 0946-8978, pp 411–416
- Egozcue JJ, Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. *Math Geol* 37:795–828
- Egozcue JJ, Pawlowsky-Glahn V (2011) Basic concepts and procedures. In: Pawlowsky-Glahn V, Buccianti A (eds) *Compositional data analysis: theory and applications*, Wiley, ISBN 978-0-470-71135-4, pp 12–28
- Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. *Math Geol* 35(3):279–300, ISSN 0882-8121
- Egozcue JJ, Daunis-i-Estadella J, Pawlowsky-Glahn V, Hron K, Filzmoser P (2012) Simplicial regression. The normal model. *J Appl Probab Stat* 6(1&2):87–108
- Egozcue J, Lovell D, Pawlowsky-Glahn V (2013) Regression between compositional data sets. In: Hron K, Filzmoser P, Templ M (eds) *Proceedings of the 5th international workshop on compositional data analysis*, Vorau
- Filzmoser P, Hron K, Templ M (2018) *Applied compositional data analysis: with worked examples in R*. Springer, Cham
- Fišerová E, Donevska S, Hron K, Bábek O, Vaňkátová K (2016) Practical aspects of log-ratio coordinate representations in regression with compositional response. *Meas Sci Rev* 16(5):235–243
- Graffelman J, van Eeuwijk F (2005) Calibration of multivariate scatter plots for exploratory analysis of relations within and between sets of variables in genomic research. *Biom J* 47(6):863–879
- Greenacre M (2019) Variable selection in compositional data using pairwise logratios. *Math Geosci* 51:649–682
- Hampel F, Ronchetti E, Rousseeuw P, Stahel W (1986) *Robust statistics. The approach based on influence functions*. Wiley, New York
- Hron K, Donevska S, Fišerová E, Filzmoser P (2013) Covariance-based variable selection for compositional data. *Math Geosci* 45(4):487–498
- Hrůzová K, Todorov V, Hron K, Filzmoser P (2016) Classical and robust orthogonal regression between parts of compositional data. *Stat A J Theor Appl Stat* 50(6):1261–1275
- Johnson R, Wichern D (2007) *Applied multivariate statistical analysis*, 6th edn. Prentice Hall, New York, p 800
- Maechler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibián-Barrera M, Verbeke T, Koller M, Conceicao ELT, Anna di Palma M (2018) *robustbase: basic robust statistics*. R package version 0.93-3
- Maronna R, Martin R, Yohai V (2006) *Robust statistics: theory and methods*. Wiley, New York
- Mateu-Figueras G, Pawlowsky-Glahn V (2008) A critical approach to probability laws in geochemistry. *Math Geosci* 40(5):489–502
- Mert C, Filzmoser P, Hron K (2015) Sparse principal balances. *Stat Model* 15(2):159–174
- Mood AM, Graybill FA, Boes DC (1974) *Introduction to the theory of statistics*, 3rd edn. McGraw-Hill, New York
- Pawlowsky-Glahn V (2003) Statistical modelling on coordinates. In: Thió-Henestrosa S, Martín-Fernández JA (eds) *Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop*, Girona (E). Universitat de Girona, ISBN 84-8458-111-X, <http://ima.udg.es/Activitats/CoDaWork2003/>
- Pawlowsky-Glahn V, Egozcue JJ (2001a) Geometric approach to statistical analysis on the simplex. *Stoch Environ Res Risk Assess (SERRA)* 15(5):384–398
- Pawlowsky-Glahn V, Egozcue JJ (2001b) Geometric approach to statistical analysis on the simplex. *Stoch Environ Res Risk Assess (SERRA)* 15(5):384–398
- Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2011) Principal balances. In: Egozcue JJ, Tolosana-Delgado R, Ortego MI (eds) *Proceedings of the 4th international workshop on compositional data analysis (2011)*, CIMNE, Barcelona, Spain, ISBN 978-84-87867-76-7
- Pawlowsky-Glahn V, Egozcue J, Tolosana-Delgado R (2015) *Modeling and analysis of compositional data*. Wiley, Chichester
- R Development Core Team (2019) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna
- Reimann C, Birke M, Demetriades A, Filzmoser P, O'Connor P (eds) (2014a) Chemistry of Europe's agricultural soils—part A: methodology and interpretation of the GEMAS data set. *Geologisches Jahrbuch (Reihe B 102)*. Schweizerbarth, Hannover

-
- Reimann C, Birke M, Demetriades A, Filzmoser P, O'Connor P (eds) (2014b) Chemistry of Europe's agricultural soils—part B: general background information and further analysis of the GEMAS data set. *Geologisches Jahrbuch (Reihe B 103)*. Schweizerbarth, Hannover
- Rousseeuw PJ, van Zomeren BC (1990) Unmasking multivariate outliers and leverage points. *J Am Stat Assoc* 8:633–639
- Salibian-Barrera M, Zamar R (2002) Bootstrapping robust estimates of regression. *Ann Stat* 30:556–582
- Salibian-Barrera M, Van Aelst S, Willems G (2008) Fast and robust bootstrap. *Stat Methods Appl* 17:41–71
- Simonoff Jeffrey S (2003) *Analyzing categorical data*. Springer, Berlin
- Tolosana-Delgado R, van den Boogaart K (2011) Linear models with compositions in R. In: Pawlowsky-Glahn V, Buccianti A (eds) *Compositional data analysis: theory and applications*. Wiley, Chichester, pp 356–371
- Van Aelst S, Willems G (2013) Fast and robust bootstrap for multivariate inference: the R package FRB. *J Stat Softw* 53(3):1–32
- van den Boogaart K, Tolosana-Delgado R (2013) *Analyzing compositional data with R*. Springer, Heidelberg
- Yohai V (1987) High breakdown-point and high efficiency estimates for regression. *Ann Stat* 15:642–665