Journal of Applied Probability and Statistics
Vol. 6, No. 1&2, pp. 87-108
ISOSS Publications 2012

SIMPLICIAL REGRESSION. THE NORMAL MODEL

Juan José Egozcue
Dept. Matemàtica Aplicada III, U. Politècnica de Catalunya (UPC), Barcelona, Spain.
Email: juan.jose.egozcue@upc.edu

Josep Daunis-i-Estadella
Dept. Informàtica i Matemàtica Aplicada, U. de Girona (UdG), Girona, Spain.
Email: pepus@ima.udg.edu

Vera Pawlowsky-Glahn
Dept. Informàtica i Matemàtica Aplicada, U. de Girona (UdG), Girona, Spain.
Email: vera.pawlowsky@udg.edu

Karel Hron
Dept. of Mathematical Analysis and Application of Mathematics, Palacký U. (UPOl), Olomouc, Czech Republic.
Email: hronk@seznam.cz

Peter Filzmoser
Dept. of Statistics and Probability Theory, Vienna U. of Technology (TU-Wien, Vienna, Austria.
Email: P.Filzmoser@tuwien.ac.at

summary

Regression models with compositional response have been studied from the beginning of the log-ratio approach for analysing compositional data. These early approaches suggested the statistical hypothesis of logistic-normality of the compositional residuals to test the model and its coefficients. Also, the Dirichlet distribution has been proposed as an alternative model for compositional residuals, but it leads to restrictive and not easy-to-use regressions. Recent advances on the Euclidean geometry of the simplex and on the logistic-normal distribution allow re-formulating simplicial regression with logistic-normal residuals. Estimation of the model is presented as a least-squares problem in the simplex and is formulated in terms of orthonormal coordinates. This estimation decomposes into simple linear regression models which can be assessed independently. Marginal normality of the coordinate-residuals suffices to check influence of covariables using standard regression tests. Examples illustrate the proposed procedures.

Keywords and phrases: Aitchison geometry, normal distribution on the simplex, isometric log-ratio transformation (ilr), orthonormal coordinates, log-ratio analysis.
Subject Classification: 62J05, 62J02, 86A32, 91B42.

1 Introduction

Compositional data appear frequently in statistical analysis. They quantitatively represent the parts of a whole and only the proportions of their parts are assumed informative. Typical examples are a chemical composition, the proportions of large counts in surveying, the structure of a stock portfolio, the distribution of household expenditures and incomes, etc. As a consequence, compositional data also occur as responses in regression models. Regression models for compositional data were first discussed in [18, 10]. In Aitchison and Shen [18] a discussion on the distribution of the residuals of the regression is enlightening. One obvious candidate was the Dirichlet distribution. The competing model was the logistic-normal family of distributions. It was shown that the Dirichlet family can be approximated by the logistic-normal distribution and thus approximately included in the logistic-normal family. Moreover, the Dirichlet family seemed to the authors too restrictive for an effective and practical use in applications [12]. Most of the material about regression with compositional responses and the distributions appropriate for residuals presented in these references keep their validity, and only a little bit about techniques can be added. However, over almost the last three decades these results have not been taken into account, and a lot of studies on Dirichlet regression for compositional responses have appeared. Recent examples are [32, 33, 2].

Recent developments on the simplex geometry [4, 35, 14, 25, 22, 20] allow to express the regression model in coordinates and to estimate its coefficients using ordinary least squares [19]. When the normal model is assumed for the residuals, its distribution is identified with the logistic-normal or additive-logistic-normal [9, 17]. In this simple case, the least squares approach can be applied to simplicial coordinates of the compositional response, and it corresponds to the maximum likelihood estimation of the model. Our objective is to present the linear regression model for compositional response in its coordinate version. The model can be estimated using ordinary least squares. Under normality of the coordinate residuals, standard statistical techniques of multiple regression can be applied. As a consequence, the logistic-normal linear regression for compositional responses is the simplest regression method, competing with other approaches like e.g. models with Dirichlet distributed residuals. Model selection is not treated here globally, but separately for each coordinate. Standard techniques in regression analysis can be used on coordinates. Also more specific techniques dealing with missing data and rounded zeros have been recently developed [31].

2 Aitchison simplicial geometry

Geometry

Compositional data of D parts are identified with equivalence classes of proportional vectors with positive components. A representative of these equivalence classes can be taken to be in the simplex of D parts (equivalently the $(D-1)$-dimensional simplex), denoted $\mathcal{S}^D$. The simplex $\mathcal{S}^D$ can be defined as the set of real vectors of D positive components adding to a constant, here assumed to be unity. If $\mathbf{x}$ is a D-vector of positive components, denote $\mathcal{C}\mathbf{x}$ its representative in the simplex. $\mathcal{C}\mathbf{x}$ is readily obtained dividing each component by their total sum, and is called the *closure of* $\mathbf{x}$.

A natural operation between elements of the simplex is *perturbation*, which plays the role of addition in the simplex. Multiplication by real numbers is called *powering*. Denoting transpose by $(\cdot)'$, compositions in $\mathcal{S}^D$ by $\mathbf{x} = (x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha)'$, $\mathbf{y} = (y_1, y_2, \ldots, y_D)'$, and $\alpha \in \mathbb{R}$, perturbation and powering are defined as

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1 y_1, x_2 y_2, \ldots, x_D y_D)' \ , \ \alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha)' \ , \tag{2.1}$$

respectively. The composition $\mathbf{n}$ with equal components is the neutral element for the perturbation. Perturbation and powering (2.1) define a $(D-1)$-dimensional vector space structure in the simplex $\mathcal{S}^D$. The Aitchison inner product in $\mathcal{S}^D$ is

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \sum_{i=1}^{D} (\log x_i \cdot \log y_i) - \frac{1}{D} \left(\sum_{j=1}^{D} \log x_j \right) \cdot \left(\sum_{k=1}^{D} \log y_k \right) \ . \tag{2.2}$$

The corresponding norm and distance are

$$\|\mathbf{x}\|_a = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a} \quad , \quad \mathrm{d}_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a \ , \tag{2.3}$$

where $\ominus$ represents the opposite operation of $\oplus$, i.e. $\ominus \mathbf{y} \equiv \oplus \left((-1) \odot \mathbf{y}\right)$. The metrics defined by eq. (2.2), resp. (2.3), is compatible with the operations in (2.1), so that the simplex $\left(\mathcal{S}^D, \oplus, \odot, \langle ., . \rangle_a\right)$ is a $(D-1)$-dimensional Euclidean space [14, 35, 4]. This constitutes the so-called Aitchison geometry of the simplex.

A consequence of the Euclidean structure of $\mathcal{S}^D$ is that an orthonormal basis of the space can be built, and a composition $\mathbf{x} \in \mathcal{S}^D$ can be represented by its coordinates with respect to such a basis. Let $\mathbf{x}^* = h(\mathbf{x})$ be the vector of $D-1$ real coordinates of $\mathbf{x}$. For each orthonormal basis, the coordinate function $h(\cdot)$ is an isometry between $\mathcal{S}^D$ and $\mathbb{R}^{D-1}$, called *isometric log-ratio transformation* [25]. Important properties of such an isometry are

$$h(\mathbf{x} \oplus \mathbf{y}) = h(\mathbf{x}) + h(\mathbf{y}) \ , \ h(\alpha \odot \mathbf{x}) = \alpha \cdot h(\mathbf{x}) \ , \tag{2.4}$$

and

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle h(\mathbf{x}), h(\mathbf{y}) \rangle \ , \ \|\mathbf{x}\|_a = \|h(\mathbf{x})\| \ , \ \mathrm{d}_a(\mathbf{x}, \mathbf{y}) = \mathrm{d}(h(\mathbf{x}), h(\mathbf{y})) \ , \tag{2.5}$$

where $\langle \cdot, \cdot \rangle$, $\| \cdot \|$ and $\mathrm{d}(\cdot, \cdot)$ are the ordinary Euclidean inner product, norm and distance in $\mathbb{R}^{D-1}$ respectively. This means that, whenever compositions are transformed into coordinates, the metrics and operations in the Aitchison geometry of the simplex are translated into the ordinary Euclidean metrics and operations in real space.

The choice of an orthonormal basis can be made following the methods developed in [22, 23]. They consist of defining a sequential binary partition (SBP) of the compositional vector. In a first step, the components of the composition are divided into two groups;

components in one group are marked with a +1 and components in the other group are marked with a −1; see Table 1, order 1 row. In a second and following steps, a previous group of parts is divided into two new groups and they are similarly marked with +1 and −1, while the components not involved are marked with 0; see second and following rows in Table 1. The number of steps required until each group contains a single component is

Table 1: Coding of a sequential binary partition (SBP) of a $D = 5$ compositional vector $\mathbf{x}$. Each row of the $(4, 5)$-matrix Θ indicates with +1 and −1 the components in each group of the partition at the corresponding order; 0 indicates that the component does not participate in the partition. Columns r, resp. s, are the number of +1, resp. −1, in the corresponding order partition. The balance-coordinate is made explicit in the last column.

order	x_1	x_2	x_3	x_4	x_5	r	s	balance
1	+1	−1	−1	+1	+1	3	2	$x_1^* = (6/5)^{1/2} \log \frac{(x_1x_4x_5)^{1/3}}{(x_2x_3)^{1/2}}$
2	+1	0	0	+1	−1	2	1	$x_2^* = (2/3)^{1/2} \log \frac{(x_1x_4)^{1/2}}{x_5}$
3	+1	0	0	−1	0	1	1	$x_3^* = (1/2)^{1/2} \log \frac{x_1}{x_4}$
4	0	−1	+1	0	0	1	1	$x_4^* = (1/2)^{1/2} \log \frac{x_3}{x_2}$

exactly $D - 1$, i.e. the dimension of $\mathcal{S}^D$. Let $\Theta = [\theta_{ij}]$ be a $(D - 1) \times D$ matrix containing the codes represented in Table 1. An element of an orthonormal basis of $\mathcal{S}^D$, and the corresponding coordinate, are associated with each row of Θ. First, for the ith-row of Θ compute the number of +1 and −1 and denote them by r_i and s_i, respectively. Then, construct the $(D - 1) \times D$ matrix $\Psi = [\psi_{ij}]$ where

$$\psi_{ij} = \theta_{ij} \, \frac{s_i^{(\theta_{ij}-1)/2}}{r_i^{(\theta_{ij}+1)/2}} \sqrt{\frac{r_i s_i}{r_i + s_i}} \ , \ i = 1, 2, \ldots, D - 1 \ , \ j = 1, 2, \ldots, D \ . \tag{2.6}$$

The matrix Ψ (2.6) has some remarkable properties, similar to those of Helmert matrices [30]. The coordinate associated with the i-th row of Θ is

$$x_i^* = \sqrt{\frac{r_i s_i}{r_i + s_i}} \ \log \frac{\prod_+ x_j^{1/r_i}}{\prod_- x_k^{1/s_i}} \ , \tag{2.7}$$

where the product subscripted + (resp. −) runs over the components marked with +1 (resp. −1) in the i-th row of Θ. The transformation into coordinates (2.7) is called isometric log-ratio transformation (ilr) [25, 22]. The coordinates are also called balances because of their particular form as ratios of geometric means of components grouped as coded in the SBP, as shown in (2.7). The computation of the balances or coordinates of the composition can be written as

$$\mathbf{x}^* = h(\mathbf{x}) = \Psi \cdot \log \mathbf{x} \ , \tag{2.8}$$

where the logarithmic function applies componentwise and the dot denotes matrix product. A composition can be readily recovered from its coordinates using the inverse ilr transformation

$$\mathbf{x} = h^{-1}(\mathbf{x}^*) = \mathcal{C}\exp(\Psi' \cdot \mathbf{x}^*) \ , \tag{2.9}$$

where exp(.) applies componentwise to the argument vector [30].

There are other ways of representing elements of the simplex. Two of them, called alr and clr [12], additive log-ratio and centered log-ratio transformations respectively, are historically previous to orthogonal coordinates, ilr, and have been used extensively. The alr transformation of a composition $\mathbf{x} \in \mathcal{S}^D$ is defined as the $(D-1)$-real vector

$$\mathrm{alr}(\mathbf{x}) = \log\left(\frac{x_1}{x_D}, \frac{x_2}{x_D}, \ldots, \frac{x_{D-1}}{x_D}\right)' \ , \tag{2.10}$$

with inverse transformation

$$\mathrm{alr}^{-1}(\mathbf{y}) = \mathcal{C}\exp\left(y_1, y_2, \ldots, y_{D-1}, 0\right)' \ , \tag{2.11}$$

where $\mathbf{y} = \mathrm{alr}(\mathbf{x}) \in \mathbb{R}^{D-1}$. The components of $\mathrm{alr}(\mathbf{x})$ are coordinates of the composition with respect to an oblique basis of the simplex [22]. This means that it can be useful for representations where the properties of $\mathcal{S}^D$ as a vector space play the main role. However, the alr representation may be not easy to use when dealing with metric properties of $\mathcal{S}^D$.

For $\mathbf{x} \in \mathcal{S}^D$, the centered log-ratio transformation clr is defined as

$$\mathrm{clr}(\mathbf{x}) = \log\left(\frac{x_1}{g(\mathbf{x})}, \frac{x_2}{g(\mathbf{x})}, \ldots, \frac{x_D}{g(\mathbf{x})}\right)' \ , \tag{2.12}$$

where $g(\cdot)$ is the geometric mean of the components of the argument. The clr representation is an isometry between $\mathcal{S}^D$ with the Aitchison geometry and the $(D-1)$-dimensional subspace of $\mathbb{R}^D$ of vectors whose components add to zero. Therefore, components of the clr transformed vectors add to zero, thus constraining its components. The clr components (2.12) permit the reconstruction of the corresponding composition

$$\mathbf{x} = \mathcal{C}\exp(\mathbf{y}) \ , \tag{2.13}$$

where $\mathbf{y} = \mathrm{clr}(\mathbf{x}) \in \mathbb{R}^D$. The clr representation of compositions is very useful to compute operations and metrics in $\mathcal{S}^D$, although a redundant component is used in the storage and in computation. Examples of use of the clr (2.12), (2.13) are the computation of compositional principal components [11, 12] and compositional biplots [15].

Elements of simplicial statistics

When dealing with random compositions, i.e. random vectors whose sample space is $\mathcal{S}^D$, the Aitchison simplicial geometry influences some elementary concepts, specially those related with the underlying metrics of the sample space. The mean and variance, and the respective estimators, are here addressed. Also the normal distribution in the simplex and its representation is briefly presented.

The concept of centre of a random composition, $\mathbf{X}$, was introduced in [13]. It can be defined as

$$\mathrm{Cen}[\mathbf{X}] = h^{-1}\mathrm{E}[h(\mathbf{X})] = \mathcal{C}\exp(\mathrm{E}[\log \mathbf{X}]) \ , \tag{2.14}$$

where $h(\cdot)$ is the coordinate function for a chosen basis in $\mathcal{S}^D$ and $\mathrm{E}[\cdot]$ is the ordinary expectation in the real space $\mathbb{R}^D$. The second member in (2.14) corresponds to a DeFinetti *gamma*-mean [1]. The third member in (2.14) is the expression given by Aitchison, which is proportional to a geometric mean. Note that the definition does not depend on the chosen basis in $\mathcal{S}^D$. The center can also be defined as the element in $\mathcal{S}^D$ minimizing the Aitchison-metric variability of $\mathbf{X}$, which does not depend on the basis [35]. In a more general framework, this definition is in agreement with the general theory developed in [26]. Given a random sample of $\mathbf{X}$, the natural estimator of $\mathrm{Cen}[\mathbf{X}]$ is the simplex-average or geometric mean [36]

$$\overline{\mathbf{X}} = \frac{1}{n} \odot \bigoplus_{i=1}^{n} \mathbf{x}_i = \mathcal{C}\left(\left(\prod_{i=1}^{n} x_{i1}\right)^{1/n}, \left(\prod_{i=1}^{n} x_{i2}\right)^{1/n}, \ldots, \left(\prod_{i=1}^{n} x_{iD}\right)^{1/n} \right)' \ , \tag{2.15}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iD})'$ is the ith-sample composition. This estimator is unbiased in the simplex, i.e. $\mathrm{Cen}[\overline{\mathbf{X}} \ominus \mathrm{Cen}[\mathbf{X}]] = \mathbf{n}$.

The metric or total variance of a random composition [13, 35] is defined in a natural way as

$$\mathrm{MVar}[\mathbf{X}] = \mathrm{E}[\mathrm{d}_a^2(\mathbf{X}, \mathrm{Cen}[\mathbf{X}])] \ . \tag{2.16}$$

There are a number of expressions of (2.16) in terms of log-ratios of the components of the random composition. When using coordinates $\mathbf{X}^*$ of the random composition with respect to a chosen basis, $\mathrm{MVar}[\mathbf{X}]$ is decomposed into variances of the coordinates [24], i.e.

$$\mathrm{MVar}[\mathbf{X}] = \sum_{j=1}^{D-1} \mathrm{Var}[X_j^*] \ , \tag{2.17}$$

where X_j^* denotes de jth-coordinate of the random composition $\mathbf{X}$. The decomposition (2.17) holds after the decomposition of the Aitchison-distance using orthonormal coordinates [22]. The estimation is then reduced to the estimation of the variances of the coordinates $\mathrm{Var}[X_j^*]$. The CoDa-dendrogram can be used for a visualization of the variance decomposition [34, 37, 24]. The covariances between coordinates complete the second order description of the variability of the random composition. They can be arranged in the variance-covariance $(D-1, D-1)$-matrix Σ whose ij-entry is $\mathrm{Cov}[X_i^*, X_j^*]$. The matrix Σ depends on the selected basis. However, the covariance endomorphism represented by Σ is invariant under changes of basis in $\mathcal{S}^D$ [26, 29].

3 Least squares regression with a compositional response.

Consider a n-sample data set in which the i-th record is made of a compositional response $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iD})'$ in $\mathcal{S}^D$, and the values of r covariates arranged in a vector $\mathbf{t}_i = (t_0, t_{i1}, t_{i2}, \ldots, t_{ir})'$, where $t_0 = 1$ is equal for each record. A prediction in the simplex $\mathcal{S}^D$ consists of a deterministic function of the covariates, also called predictor, $\mathbf{p}(\mathbf{t}) \in \mathcal{S}^D$; and a perturbation-additive error or residual $\mathbf{e} \in \mathcal{S}^D$. A linear predictor in the simplex is

$$\mathbf{p}(\mathbf{t}) = \bigoplus_{k=0}^{r} (t_k \odot \mathbf{b}_k) \ , \tag{3.1}$$

where the coefficients $\mathbf{b}_k \in \mathcal{S}^D$. The predictor (3.1), is a linear combination of compositional coefficients $\mathbf{b}_k$, with respect to the Aitchison geometry of the simplex, where the coefficients of the combination are the real covariates. The covariate $t_0 = 1$ provides a constant term in the predictor.

The least squares regression problem is to find estimates, $\widehat{\mathbf{b}}_k$, of the compositional coefficients $\mathbf{b}_k$, $k = 0, 1, \ldots, r$, in

$$\mathbf{x}_i = \mathbf{b}_0 \oplus \bigoplus_{k=1}^{r} (t_{ik} \odot \mathbf{b}_k) \oplus \mathbf{e}_i \ , \ i = 1, 2, \ldots, n \ , \tag{3.2}$$

minimizing the sum of square-norms of the error

$$\mathrm{SSE} = \sum_{i=1}^{n} \|\mathbf{e}_i\|_a^2 = \sum_{i=1}^{n} \|\mathbf{p}(\mathbf{t}_i) \ominus \mathbf{x}_i\|_a^2 \ . \tag{3.3}$$

The regression model (3.2) contains $(r+1)\times D$ parameter values to be determined. However, the $\mathbf{b}_k$'s are in the simplex and $D-1$ components determine these coefficients and, therefore, there are only $(r+1) \times (D-1)$ parameters to be estimated from the data. It is worth to remark that all familiar geometrical concepts in (3.2) and (3.3), like linearity, deviation, norm, are here referred to the Aitchison geometry of the simplex. Accordingly, SSE (3.3) cannot be compared to similar expressions in which the norms and operations are those of the standard real Euclidean space. The adequacy of SSE as a target function to be minimized relays on the compositional character of the response and the consequent measurement of deviations in $\mathcal{S}^D$.

Assume that the least-squares estimate of the compositional coefficients are $\widehat{\mathbf{b}}_k$, thus defining the predictor $\widehat{\mathbf{p}}(\mathbf{t})$. The corresponding estimated residuals are $\widehat{\mathbf{e}}_i$ and $\widehat{\mathrm{SSE}}$ denotes the minimized sum of squares. Similarly to the standard multiple linear regression analysis, the total sum of squares $\widehat{\mathrm{SST}}$, defined as

$$\widehat{\mathrm{SST}} = \sum_{i=1}^{n} \|\mathbf{x}_i \ominus \overline{\mathbf{X}}\|_a^2 \ , \tag{3.4}$$

is considered. The statistics $\overline{\mathbf{X}}$ in (3.4) is the geometric average of the sample response as defined in (2.15). The statistics $n^{-1} \cdot \widehat{\mathrm{SST}}$ is an estimator of the total variance of the responses MVar[$\mathbf{X}$] (2.16). Also, a sum of squares explained by the regression model can be defined as

$$\widehat{\mathrm{SSR}} = \sum_{i=1}^{n} \|\widehat{\mathbf{p}}(\mathbf{t}_i) \ominus \overline{\mathbf{X}}\|_a^2 \ , \tag{3.5}$$

which gives rise to a decomposition of $\widehat{\mathrm{SST}}$:

$$\widehat{\mathrm{SST}} = \widehat{\mathrm{SSR}} + \widehat{\mathrm{SSE}} \ . \tag{3.6}$$

The reasoning to arrive to the decomposition (3.6) is parallel to that of the ordinary real multivariate linear regression. Similarly, a determination coefficient of the regression model can be defined as

$$R^2 = \frac{\widehat{\mathrm{SSR}}}{\widehat{\mathrm{SST}}} = 1 - \frac{\widehat{\mathrm{SSE}}}{\widehat{\mathrm{SST}}} \ , \tag{3.7}$$

which is interpreted as the per unit of metric-variance of the compositional response explained by the regression.

The least-squares problem can be efficiently solved expressing the compositional responses in coordinates, specifically with respect to an orthonormal basis of the simplex. If $h(\cdot)$ is the coordinate function for the chosen orthonormal basis, denote $\mathbf{x}_i^* = h(\mathbf{x}_i)$, $\mathbf{e}_i^* = h(\mathbf{e}_i)$ for $i = 1, 2, \ldots, n$; and $\mathbf{b}_k^* = h(\mathbf{b}_k)$, $k = 0, 1, \ldots, r$. Taking coordinates in (3.2), the transformed model is

$$\mathbf{x}_i^* = \mathbf{b}_0^* + \sum_{k=1}^{r} (t_{ik} \cdot \mathbf{b}_k^*) + \mathbf{e}_i^* \ , \ i = 1, 2, \ldots, n \ , \tag{3.8}$$

and, using (2.17),

$$\mathrm{SSE} = \sum_{i=1}^{n} \|\mathbf{e}_i^*\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{D-1} (e_{ij}^*)^2 \ . \tag{3.9}$$

Eq. (3.9) is a consequence of the isometric character of $h(\cdot)$: the Aitchison norm of a composition is equal to the ordinary real Euclidean norm of its coordinates (2.5). In the expression of SSE (3.9), the order of the sums can be inverted and, being all terms non-negative, the minimization of SSE in coordinates is equivalent to the separate minimization of the $D-1$ terms

$$\mathrm{SSE}_j = \sum_{i=1}^{n} (e_{ij}^*)^2 = \sum_{i=1}^{n} \left(x_{ij} - \sum_{k=0}^{r} t_k b_{kj}^* \right)^2 \ , \ j = 1, 2, \ldots, D-1 \ , \tag{3.10}$$

where b_{kj}^* is the j-th coordinate of the compositional coefficient $\mathbf{b}_k$. Comparing (3.9) and (3.10), the Pythagorean decomposition $\sum_{j=1}^{D-1} \mathrm{SSE}_j = \mathrm{SSE}$ is easily obtained. For the j-th coordinate, (3.10) implies the ordinary least-squares solution of the real regression model

$$x_{ij}^* = \sum_{k=0}^{r} t_k b_{kj}^* + e_{ij}^* \ , \ i = 1, 2, \ldots, n \ , \tag{3.11}$$

where e^*_{ij} is the j-th coordinate of the compositional residual $\mathbf{e}_i$. Eqs. (3.10) and (3.11) imply that the least-squares regression problem in the simplex (3.2), (3.3) is equivalent to $D-1$ ordinary least-squares problems for the coordinates (3.10) and (3.11). Remarkably, the least-squares problems for the coordinates can be solved independently. Moreover, the results are independent of the selected orthonormal basis: although the coordinates of the obtained coefficients $\mathbf{b}_k$ and residuals $\mathbf{e}_i$ depend on the selected basis, the reconstructed compositional coefficients and residuals using (2.9) do not.

For each regression problem (3.11), (3.10), the sum of squares decomposition holds, i.e. $\widehat{\mathrm{SSE}} = \sum_{j=1}^{D-1} \widehat{\mathrm{SSE}}_j$ and $\widehat{\mathrm{SSR}} = \sum_{j=1}^{D-1} \widehat{\mathrm{SSR}}_j$. The determination coefficient can also be expressed in terms of the sums of squares of the regression for the coordinates,

$$R^2 = \frac{\sum_{j=1}^{D-1} \widehat{\mathrm{SSR}_j}}{\widehat{\mathrm{SST}}} = \frac{\sum_{j=1}^{D-1} \widehat{\mathrm{SST}}_j \cdot R_j^2}{\widehat{\mathrm{SST}}} \ , \tag{3.12}$$

where $R_j^2 = \widehat{SSR_j}/\widehat{SST_j}$ is the determination coefficient for the regression of the jth-coordinate of the response.

The whole procedure may be summarized in the following steps: (i) select an orthonormal basis, possibly using a sequential binary partition (SBP) of the compositional response vector; (ii) represent the compositional response by means of its orthonormal coordinates, possibly balance-coordinates; (iii) perform the least-squares estimation of the regression coefficients and the sums of squares for each coordinate of the response using the available covariates; (iv) reconstruct, if necessary, the compositional coefficients, predictor and residuals. These steps correspond to the *principle of working on coordinates* [8].

The standard practice in logistic regression [16, 10, 5] , in spatial cokriging [38] or even in simplicial regression [4, 19], has not been to use the ilr transformation (orthonormal basis representation) but the alr transformation (oblique basis representation). A natural question is which is the difference in the least-squares results when using these two different representations of the compositional response. In fact, there is no difference in the estimated compositional coefficients of the regression model (3.2) and, consequently, the compositional residuals are also equal. The difference appears when trying to obtain the decomposition of $\widehat{\mathrm{SST}}$ (3.6) into the alr-coordinate contributions (3.4). When using alr-coordinates, $\sum_{j=1}^{D-1} \widehat{\mathrm{SST}}_j \geq \widehat{\mathrm{SST}}$, $\sum_{j=1}^{D-1} \widehat{\mathrm{SSR}}_j \neq \widehat{\mathrm{SSR}}$, and $\sum_{j=1}^{D-1} \widehat{\mathrm{SSE}}_j \neq \widehat{\mathrm{SSE}}$. In order to compute the sums of squares it is then necessary to obtain the compositional predictors and residuals and to compute $\widehat{\mathrm{SSR}}$ and $\widehat{\mathrm{SSE}}$ using their definition (3.5),(3.3) and the Aitchison-norm (2.3). It is remarkable that in standard multinomial logistic regression there are difficulties for defining a determination coefficient. This is related to the representation of the response probabilities using alr-coordinates.

4 The normal model of compositional residuals

4.1 Normal distribution on the simplex

A statistical analysis of a regression model requires further hypotheses on the distribution of the residuals. The simplest model with compositional residuals is that of the logistic-normal distribution introduced by Aitchison and Shen [18], also to be found in [10, 12]. There, the logistic-normal model is compared with the Dirichlet distribution approach for the residuals. The main argument against the Dirichlet approach is that this distribution is too restrictive and imposes strong conditions on the dependence between components. Moreover, the Dirichlet distribution can be suitably approximated (in the sense of Kullback-Leibler divergence) by some distributions in the logistic-normal family. This gives sense to the point put forward by Aitchison and Shen [18], which remains still open: *Can we develop satisfactory tests of the separate families, Dirichlet and logistic-normal, along the lines of Cox (1962)? In particular, to what extent are current tests of multivariate normality powerful against the Dirichlet alternative?*

The main argument in favour of the logistic-normal distribution is the invariance of the family under perturbations in the simplex. An important consequence is the central limit theorem for the logistic-normal distribution, sketched in Aitchison [12]. This makes the logistic-normal distribution a natural one.

The logistic-normal distribution can be defined in different ways. The original definitions by J. Aitchison are based on the normality of the alr coordinates of a random composition. More recently, and following the lines proposed by Eaton [26], an intrinsic definition independent of coordinates is available [29]. Here the definition is based on the representation in orthonormal coordinates [9, 7, 6].

Consider a random composition $\mathbf{X} \in \mathcal{S}^D$ whose representation in coordinates with respect to a selected orthonormal basis is $\mathbf{X}^* \in \mathbb{R}^{D-1}$, $\mathbf{X}^* = h(\mathbf{X})$. The random composition $\mathbf{X}$ has a logistic-normal distribution or, equivalently, a normal distribution in the simplex, whenever $\mathbf{X}^*$ has a multivariate normal distribution, i.e. $\mathbf{X}^* \sim \mathcal{N}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$. Then $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}^D}(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, with $\mathrm{Cen}[\mathbf{X}] = h^{-1}(\boldsymbol{\mu}^*)$.

When the normal in the simplex is represented by a probability density, it is better to take the Aitchison measure than the Lebesgue measure as reference. The probability density of $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}^D}(\mu^*, \Sigma^*)$ with respect to the Aitchison measure is

$$f_{\mathbf{X}}^{\mathcal{S}}(\mathbf{x}) = (2\pi)^{-(D-1)/2} \ |\boldsymbol{\Sigma}^*|^{-1/2} \ \exp\left(-\frac{1}{2}(\mathbf{x}^* - \boldsymbol{\mu}^*)'\boldsymbol{\Sigma}^{*-1}(\mathbf{x}^* - \boldsymbol{\mu}^*)\right) \ , \tag{4.1}$$

where $\mathbf{x}$ is an element of the simplex $\mathcal{S}^D$ and $\mathbf{x}^*$ is the vector of coordinates with respect to a given orthonormal basis. Note the absence of a Jacobian in (4.1); it is cancelled when changing the reference measure [9]. The density (4.1) is actually the Radon-Nikodym derivative of the probability with respect to the Aitchison measure in the simplex.

If the Lebesgue measure is used as reference, the logistic-normal density has the expression

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{(2\pi)^{-(D-1)/2} \mid \mathbf{\Sigma}^* \mid^{-1/2}}{\sqrt{D}\; x_1 x_2 \cdots x_D} \exp\left(-\frac{1}{2}\; (\mathbf{x}^* - \boldsymbol{\mu}^*)' {\mathbf{\Sigma}^*}^{-1} (\mathbf{x}^* - \boldsymbol{\mu}^*) \right) \;, \qquad (4.2)$$

where the denominator is the Jacobian of the coordinate transformation [9].

Normal compositional residuals

The standard statistical model for linear regression assumes that the residuals are independent and normally distributed. Similarly, independence and normality in the simplex are here assumed for the compositional residuals in the regression model (3.2). This assumption permits to use likelihood ratio tests to check global hypotheses on the regression models. They were developed in [12] and then used in a lattice of hypothesis with increasing complexity, to arrive to an appropriate regression. No further development is here offered in these aspects. However, expressing the regression model in orthonormal coordinates, conveys an additional result, not clearly developed previously: the standard battery of testing hypotheses for linear regression models can be applied to the regression model for each orthogonal coordinate (3.11). Therefore, marginal normality of each coordinate residual is enough to use regression tests based on normality. However, these marginal tests depend in general on the selected basis of the simplex.

5 Illustrative examples

In the following examples, we apply the above mentioned theoretical considerations to real data cases from different fields of interest, namely economics and geochemistry. Special attention will be devoted to the construction of balances and to the interpretation of results.

Example 1 (Household expenditures) The first data set comes from Eurostat (European Union statistical information service) and represents mean consumption expenditures of households on 12 domestic year costs in all 27 Member States of the European Union (EU) in 2005; it is available at `http://epp.eurostat.ec.europa.eu/ statistics_explained/index.php /Household_consumption_expenditure`. The data are displayed in Table 2, together with the gross domestic product (GDP) for 2009, one of the well known measures of a country's overall economic performance that was obtained from public sources of the internet encyclopedia Wikipedia. The GDP represents the market value of all final goods and services made within the borders of a country in a year. In order to offer a better insight into the construction and interpretation of balances, we focus on a subcomposition of four parts, that include expenditures on foodstuff, housing (including water, electricity, gas and other fuel), health, and communications. The first two parts thus represent basic costs, while the latter two rather "external" costs that seem to be more or less related to economic status and, consequently, also to quality of life in each member state. However, to see the influence of GDP, not the absolute values as in Table 2, but the ratios between the expenditures are of interest. Since the absolute values are influenced by the overall price levels in the single states, their direct analysis would lead to meaningless results. The closed geometric mean of the chosen expenditures (denoted $x_1, \ldots, x_4$) is $\overline{\mathbf{X}} = (0.364, 0.496, 0.066, 0.074)'$, i.e. the expenditures on housing clearly dominate.

Table 2 shows that the GDP of Luxembourg is considerably higher than for the other countries. Since the least squares method is very sensitive to outlying observations, especially in the direction of an explanatory variable, this could essentially change the results of regression analysis and affect the final interpretation. For this reason, we exclude Luxembourg from further computations.

To see the effect of the GDP on both basic and external costs using regression analysis, we decompose the relative information contained in the (sub)composition, into balances. Here it seems natural to separate the parts x_1 and x_2, representing the basic costs, from the external ones, x_3 and x_4. The corresponding SBP is displayed in Table 3. Thus, the first coordinate, x_1^*, represents the balance between the parts x_1, x_2 and the parts x_3, x_4, or equivalently expressed, it explains the four ratios between foodstuff and housing on one side, and health and communications on the other side. The second balance, x_2^*, then explains the ratio between foodstuff and housing, and x_3^* the remaining ratio between health and communications. The variances of the balances are $\mathrm{Var}[X_1^*] = 0.060$, $\mathrm{Var}[X_2^*] = 0.166$ and $\mathrm{Var}[X_3^*] = 0.144$. Taking into account Eq. (2.17) for the metric variance, $\mathrm{MVar}[\mathbf{X}]$, one can conclude that the second and third balance explain most of the variability contained in the composition.

For all three balances we apply the regression model according to (3.11). The obtained regression lines are displayed in Figure 1. Since in the following we assume normal distributed residuals we have to check this assumption. For this reason, we employ the well known Quantile-Quantile (Q-Q) plot that compares theoretical quantiles of the normal distribution with the corresponding quantiles coming from the regression residuals. If the points in the plot lie approximately on a line, the residuals are approximating a normal distribution. Although here some deviations are clearly visible, see Figure 1 (lower row), the assumption of normality seems to be reasonable. This can be consequently checked also with some normality tests; e.g., with the well-known Anderson-Darling test [21] we obtain p-values 0.632, 0.409 and 0.401, respectively, meaning that the hypothesis of normal distribution can not be rejected in all three cases.

Table 4 summarizes the estimated regression coefficients, together with results from the inference statistics. From Figure 1 (upper left) it can be seen that the linear relation between the first coordinate and GDP is very poor. This means that GDP has nearly no influence on the ratios between parts from the variable groups "basic" and "external", represented by the first balance. Also the low coefficient of determination of $R_1^2 = 0.045$ confirms this finding. However, one should be careful with more general conclusions, because by construction of the first balance, a nearly constant relation of the balance to GDP can also be reached by an increase of one ratio and a decrease of the other ratio by about the same amount. For the second balance, that describes only the ratio foodstuff/housing, a decreas-

Table 2: GDP per capita (2009) and mean consumption expenditures of households on 12 domestic year costs (2005; both in Euro) in all 27 Member States of the European Union.

Member State	GDP	foodstuff	housing	alcohol and tobacco	clothing and footwear	household equipment	health	transport	communications	recreation and culture	education	restaurants and hotels	miscellaneous
Austria	29700	3933	6732	847	1682	1868	946	4863	793	3809	242	1660	2792
Belgium	28100	4043	7610	669	1425	1687	1400	3863	878	2868	136	1894	3576
Bulgaria	9200	2238	2461	269	218	213	305	355	325	204	34	255	220
Cyprus	22500	5158	7381	646	2649	2008	1624	4980	1164	2044	1354	2830	2370
Czech Republic	18900	2503	2444	347	679	815	239	1351	555	1289	66	619	1234
Denmark	27600	2872	7194	785	1168	1459	639	3331	583	2738	100	960	2233
Estonia	14300	2440	3240	300	601	568	282	1087	596	691	145	339	559
Finland	26600	3086	6614	588	934	1238	852	3818	693	2731	51	1021	2733
France	26000	3733	7339	650	1853	1693	1167	3777	914	1926	165	1277	3392
Germany	27300	3185	8445	489	1355	1543	1024	3790	828	3168	236	1212	3226
Greece	23200	4801	7442	1045	2154	1929	1824	3222	1174	1285	738	2661	2701
Hungary	14800	2413	2073	380	537	498	440	1511	696	909	90	343	803
Ireland	32200	4491	8520	2032	1851	2613	904	4203	1255	3670	687	2190	3956
Italy	23300	5359	8512	506	2013	1670	1132	3420	621	1680	202	1428	2242
Latvia	11300	3091	1810	329	778	546	394	1155	610	667	145	557	508
Lithuania	12500	3166	1776	332	743	392	445	762	435	402	102	429	393
Luxembourg	63300	4851	15611	865	3343	3702	1351	8403	1139	3869	223	4098	4478
Malta	18800	6082	2596	786	2387	3070	869	4758	837	2879	352	2030	1960
Netherlands	31200	3089	7513	625	1694	1888	371	3196	903	3193	306	1647	4945
Poland	14100	2704	3341	262	489	478	485	862	512	662	138	180	571
Portugal	18100	3243	5560	477	861	994	1264	2693	616	1182	356	2263	1359
Romania	10200	2355	832	307	333	201	205	344	259	224	45	58	162
Slovakia	16500	2910	2517	333	661	494	330	986	506	712	92	520	713
Slovenia	21200	3966	5483	575	1678	1389	356	3717	950	2234	202	1035	2220
Spain	24500	4685	7874	586	1786	1211	577	2743	701	1659	292	2414	1499
Sweden	28300	2913	8250	531	1270	1640	638	3623	791	3398	8	981	1569
United Kingdom	27600	3159	9458	753	1585	2092	383	4305	852	3943	457	2558	2415
Abbreviation	t	x_1	x_2	-	-	-	x_3	-	x_4	-	-	-	-

Table 3: Sequential binary partition (SBP) for household expenditures on 'foodstuff' (x_1), 'housing' (x_2), 'health' (x_3) and 'communications' (x_4). The balance-coordinate is made explicit in the last column.

order	x_1	x_2	x_3	x_4	r	s	balance
1	$+1$	$+1$	-1	-1	2	2	$x_1^* = \log \frac{(x_1x_2)^{1/2}}{(x_3x_4)^{1/2}}$
2	$+1$	-1	0	0	1	1	$x_2^* = (1/2)^{1/2} \log \frac{x_1}{x_2}$
3	0	0	$+1$	-1	1	1	$x_3^* = (1/2)^{1/2} \log \frac{x_3}{x_4}$

Table 4: Results of regression analysis for the first, second and third balance, respectively (see Table 3). Displayed are estimated coefficients of intercept and slope, values of the t-statistic and their corresponding p-values (under the assumption of normality).

coefficient	estimated value	t-statistic	p-value
b_{01}^*	1.648	10.435	2.13×10^{-10}
b_{11}^*	7.474×10^{-6}	1.065	0.297
b_{02}^*	0.786	4.814	6.67×10^{-5}
b_{12}^*	-4.684×10^{-5}	-6.461	1.11×10^{-6}
b_{03}^*	-0.212	-0.953	0.350
b_{13}^*	5.921×10^{-6}	0.599	0.554

ing trend is clearly visible, confirmed by the corresponding t-statistic (see Table 4) as well as by $R_2^2 = 0.635$. From the construction of the coordinate x_2^* (see Table 3), this corresponds to a decreasing ratio between foodstuff and household expenditures for increasing values of GDP. This is somewhat in contradiction with our intuition, since we would expect a rather constant relation between GDP and the ratio of the basic costs. Finally, the regression of the third balance on GDP shows that the ratio between the selected external costs is independent from the economic status of the member states; here $R_3^2 = 0.015$. Again, one would rather expect a systematic influence of the GDP. Using (3.12) we obtain the coefficient of determination for the whole regression model, $R^2 = 0.323$. Note that another choice of SBP would enable to focus also on the other ratios induced by the investigated composition.

Example 2 (Concentrations of chemical elements) Here we employ the well-known Kola data set which resulted from a large geochemical mapping project, carried out from 1992 to

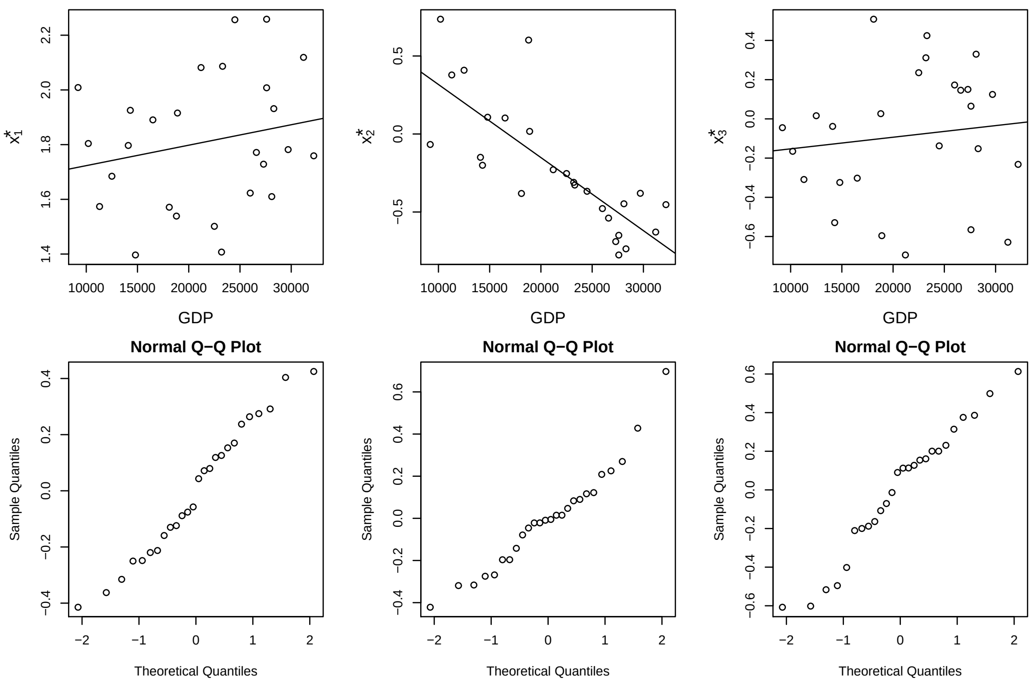

Figure 1: Regression for the first (upper left), second (upper middle) and third (upper right) balance in dependence on GDP, together with the resulting regression lines. In the lower row Q-Q plots for residuals of the corresponding regression models are displayed.

1998 by the Geological Surveys of Finland and Norway, and the Central Kola Expedition, Russia. An area covering 188000 km^2 in the Kola peninsula of Northern Europe was sampled (Figure 2). In total, approximately 600 samples of soil were taken in four different layers (moss, O-horizon, B-horizon, C-horizon) and subsequently analyzed by a number of different techniques for more than 50 chemical elements. The project was primarily designed to reveal the environmental conditions in the area; more details can be found in [3]. The whole data set is available in package `StatDA` [28] of the statistical software R. For our study, three chemical elements from the O-horizon were taken, Fe (iron, x_1), K (potassium, x_2), and P (phosphorus, x_3), and their values are reported in mg/kg. The element concentrations are depending on different geological processes, but also other effects play an important role, like the climatic zones (corresponding to the latitude) or the elevation (Figure 2). Especially elements like potassium (K) and phosphor (P) are likely to depend on latitude and/or elevation, because they both form a nutrient base for plants. However, from the maps of the single element concentrations [3] it is not easy to detect whether elevation is indeed a dominant effect for the element concentrations. With three-part compositions, we

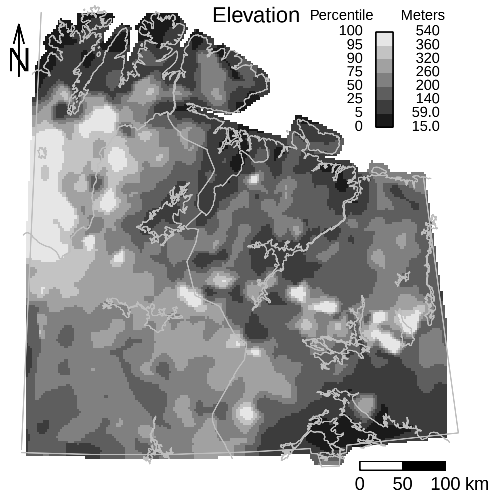

Figure 2: Map of the Kola peninsula, lighter shadings correspond to higher altitude.

have the possibility to visualize the observations in a ternary diagram (Figure 3, left). Here, the symbol size is proportional to the elevation. However, any systematic pattern is not visible (analogously also longitude and latitude as location variables would show no clear effects).

The distribution of the concentrations of Fe, and in particular of K and P in the study area can be revealed by employing the same strategy for the sequential binary partition as in the previous example. Thus, in the first balance we separate Fe from the other elements and the second balance of interest will correspond to the logratio between both nutrient base elements K and P. The sequential binary partition is displayed in Table 5 and the resulting coordinates are shown in Figure 3 (right). Here some departures from the main data cloud are clearly visible, and they are due to outliers in the ratio K and P, expressed by the second balance. One of the main questions is whether the concentrations

Table 5: Coding a sequential binary partition (SBP) for the composition Fe, K, P of the O-horizon in Kola data. The balance-coordinate is made explicit in the last column.

order	x_1	x_2	x_3	r	s	balance
1	$+1$	-1	-1	1	2	$x_1^* = (2/3)^{1/2} \log \frac{x_1}{(x_2 x_3)^{1/2}}$
2	0	-1	$+1$	1	1	$x_2^* = (1/2)^{1/2} \log \frac{x_3}{x_2}$

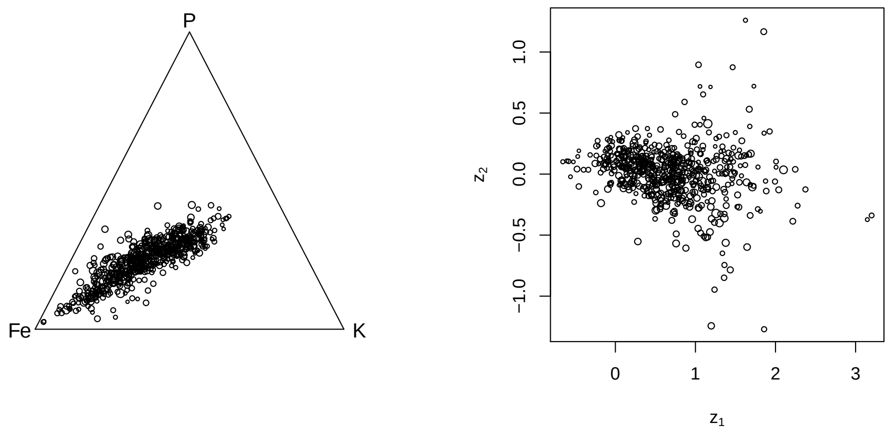

Figure 3: Ternary diagram (left) and coordinate representation (right) of the elements Fe, K, P from O-horizon of the Kola data.

of the elements are uniformly distributed in the study area, and whether an influence of elevation can be demonstrated. For this purpose, we construct regression models for both balances, with longitude, latitude and elevation as explanatory variables. The results are summarized in Table 6. The first balance, that explains the ratios Fe/K and Fe/P, confirms our preliminary expectations. Elevation is significant in the regression model, and longitude is nearly significant on the usual significance level $\alpha = 0.05$. Since Fe is supposed to be independent from location and elevation, the parts K and/or P will be responsible for the significance. Also for the ratio P to K, expressed by the second balance, both elevation and longitude play an important role in the regression model. The elevation is highly significant, with a p-value of 6.92×10^{-9}, revealing that the construction of the balances for the regression model was able to confirm our expectations that plant nutrients indeed depend on the altitude. In fact, the ratio of P to K is increasing with increasing elevation.

The Q-Q plots of the residuals for both balances are presented in Figure 4 (upper row). They show certain deviations from normality, and thus care has to be taken with the validity of the results. A possible solution could be to use robust methods that are able to deal with certain deviations from normality [27]. On the other hand, the above findings can be compared with maps of the values of both balances, see Figure 4, lower row. Indeed, the effect of elevation on the first balance is visible in the map (lower left), and even more clearly visible for the second balance (lower right).

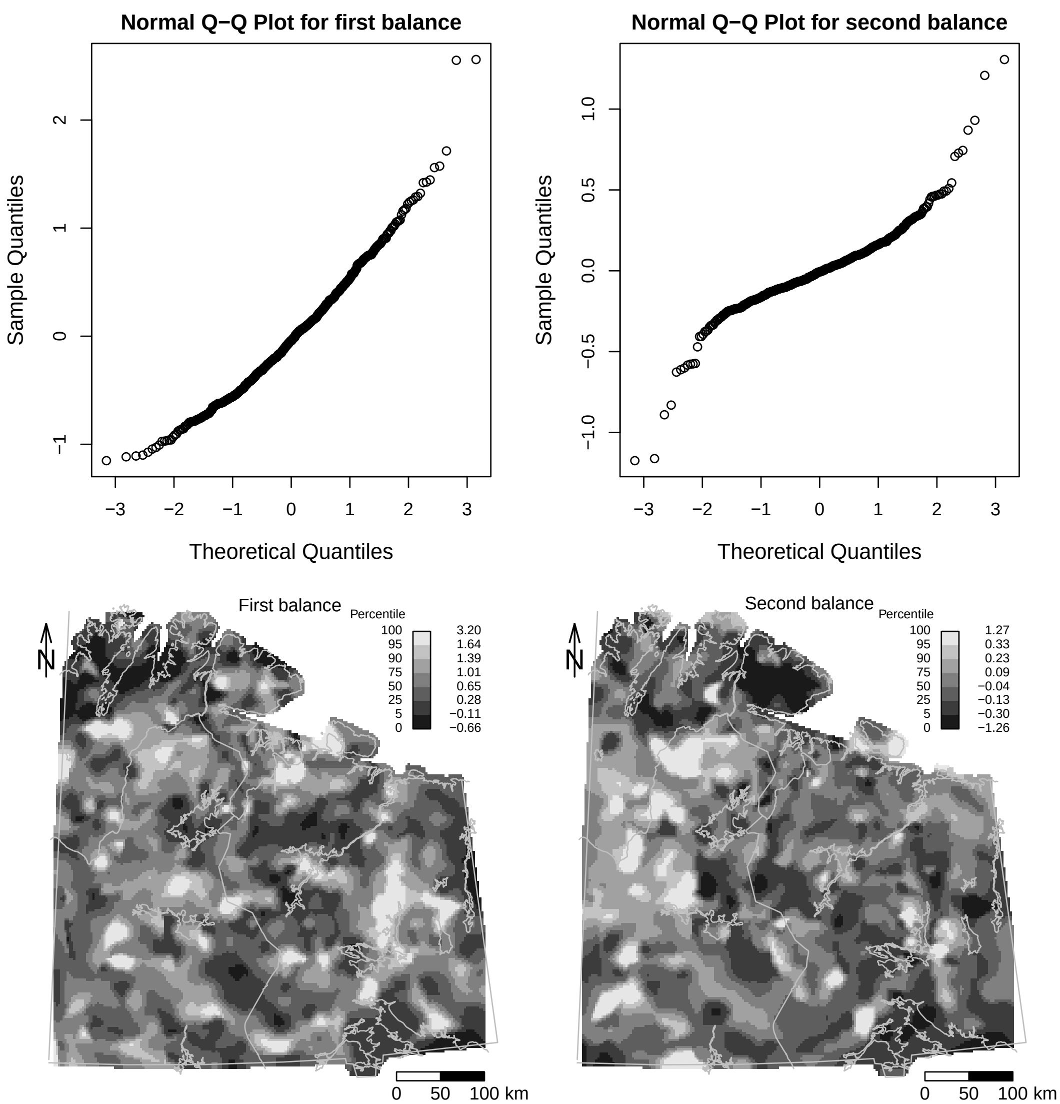

Figure 4: Q-Q plots of the residuals resulting from regressions with the first and second balance, respectively (upper row), and maps of the balances (lower row).

Table 6: Results of regression analysis for the first and second balance, respectively. Displayed are estimated coefficients of intercept and slope, values of the t-statistic and their corresponding p-values (under the assumption of normality).

parameter	coefficient	estimated value	t-statistic	p-value
intercept	b_{01}^*	2.431	1.616	0.107
longitude	b_{11}^*	3.290×10^{-7}	1.752	0.080
latitude	b_{21}^*	-2.717×10^{-7}	-1.424	0.155
elevation	b_{31}^*	5.502×10^{-4}	2.144	0.032
intercept	b_{02}^*	0.374	0.621	0.5350
longitude	b_{12}^*	-1.358×10^{-7}	-1.806	0.0715
latitude	b_{22}^*	-5.685×10^{-8}	-0.744	0.4570
elevation	b_{32}^*	6.035×10^{-4}	5.875	6.92×10^{-9}

6 Conclusion

Regression models with compositional response were proposed in the eighties. The natural statistical hypothesis was that compositional residuals follow logistic-normal distribution. Using the Euclidean structure of the simplex, the response variables can be represented using orthogonal coordinates. The estimation of model coefficients is formulated as a least-squares problem with respect to the Aitchison geometry of the simplex and then translated into coordinates. Each coordinate can be studied separately under marginal normality of coordinate residuals using a standard and simple regression model. Formulated in this way, simplicial regression under logistic-normal residuals is a natural and easy-to-use model.

References

[1] B. De Finetti, 1990. Theory of Probability. A critical introductory treatment. Wiley Classics Library (First published Wiley & Sons, 1974), Vol. 1 and 2. 300p.

[2] B. Smith, and W. Rayens, 2002. Conditional generalized Liuville distributions on the simplex statistics. Statistics, 36, Number 2, 185-194.

[3] C., M. Reimann, Äyräs, V. Chekushin, I. Bogatyrev, R. Boyd, P. d. Caritat, R. Dutter, T. Finne, J. Halleraker, O. Jæger, G. Kashulina, O. Lehto, H. Niskavaara, V. Pavlov, M. Räisänen, T. Strand, and T. Volden, 1998: Environmental geochemical atlas of the Central Barents Region. Geological Survey of Norway (NGU), Geological Survey of

Finland (GTK), and Central Kola Expedition (CKE), Special Publication, Trondheim, Espoo, Monchegorsk. 745 p.

[4] D. Billheimer, P. Guttorp, and W. Fagan, 2001. Statistical interpretation of species composition. Journal of the American Statistical Association, 96, Number 456, 1205-1214.

[5] D. McFadden, 1974. Frontiers in Econometrics. Academic Press, 105-142.

[6] G. Mateu-Figueras and V. Pawlowsky-Glahn, 2008. A critical approach to probability laws in geochemistry. Mathematical Geosciences, 40, Number 5, 489-502.

[7] G. Mateu-Figueras, V. Pawlowsky-Glahn and C. Barceló-Vidal, 2005. The additive logistic skew-normal distribution on the simplex. Stoch. Environ. Res. Risk Assess., 19, 205-214.

[8] G. Mateu-Figueras, V. Pawlowsky-Glahn, and J. J. Egozcue, 2011. The principle of working on coordinates. In: Pawlowsky-Glahn, V. and Buccianti A. (eds.), Compositional Data Analysis: Theory and Applications, Wiley, Chichester UK, (isbn 0-470-71135-3), ch. 3, 31-41.

[9] G. Mateu-Figueras, 2003. Models de distribució sobre el símplex. PhD thesis, Universitat Politècnica de Catalunya, Barcelona, Spain, 202.

[10] J. Aitchison 1982.The statistical analysis of compositional data (with discussion). Journal of the Royal Statistical Society, Series B (Statistical Methodology). 44, Number 2, 139-177.

[11] J. Aitchison 1983. Principal component analysis of compositional data. Biometrika. 70, Number 1, 57-65.

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

[13] J. Aitchison 1997. The one-hour course in compositional data analysis or compositional data analysis is simple. In Proceedings of IAMG'97 — The third annual conference of the International Association for Mathematical Geology, Pawlowsky-Glahn, V., editor, volume I, II and addendum, International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E), 1100 p, 3-35.

[14] J. Aitchison, C. Barceló-Vidal, J. J. Egozcue, and V. Pawlowsky-Glahn, 2002. A concise guide for the algebraic-geometric structure of the simplex, the sample space for compositional data analysis. Proceedings of IAMG'02 — The eigth annual conference of the International Association for Mathematical Geology, Bayer, U., Burger, H., and Skala, W., editors, volume I and II, International Association for Mathematical Geology, Selbstverlag der Alfred-Wegener-Stiftung, Berlin, 387-392.

[15] J. Aitchison, and M. Greenacre, 2002. Biplots for compositional data. Journal of the Royal Statistical Society, Series C (Applied Statistics), 51, Number 4, 375-392.

[16] J. Aitchison, J. W. Kay, and I. J. Lauder, 2005. Statistical concepts and applications in clinical medicine. Chapman and Hall/CRC, London (UK). 339p.

[17] J. Aitchison, G. Mateu-Figueras, and K. W. Ng, 2004. Characterisation of distributional forms for compositional data and associated distributional tests. Mathematical Geology. 35, Number 6, 667-680.

[18] J. Aitchison and S. M. Shen, 1980. Logistic-normal distributions. Some properties and uses. Biometrika. 67, Number 2, 261-272.

[19] J. Daunis-i-Estadella, J. J. Egozcue, and V. Pawlowsky-Glahn, 2002. Least squares regression in the simplex. Proceedings of IAMG'02 — The eigth annual conference of the International Association for Mathematical Geology, Bayer, U., Burger, H., and Skala, W., editors, volume I and II, International Association for Mathematical Geology, Selbstverlag der Alfred-Wegener-Stiftung, Berlin. 411-416.

[20] J. J. Egozcue, Barceló-Vidal, C., Martín-Fernández, J. A., Jarauta-Bragulat, E., Díaz-Barrero, J. L. and Mateu-Figueras, G., 2011. Elements of simplicial linear algebra and geometry. In: Pawlowsky-Glahn, V. and Buccianti A. (eds.), Compositional Data Analysis: Theory and Applications, Wiley, Chichester UK, (isbn 0-470-71135-3), ch. 11, 141-146.

[21] T. Anderson and D. Darling, 1952. Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes. Annals of Mathematical Statistics. 23, Number 2, 193-212.

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

[23] J. J. Egozcue, and V. Pawlowsky-Glahn, 2006. Simplicial geometry for compositional data. In: Buccianti, A., Mateu-Figueras, G. and Pawlowsky-Glahn, V., (eds) Compositional Data Analysis in the Geosciences: From Theory to Practice. Geological Society, London, (isbn 1-86239-205-6), 145-159.

[24] J. J. Egozcue, and V. Pawlowsky-Glahn, 2011. Basic concepts and procedures. In: Pawlowsky-Glahn, V. and Buccianti A. (eds.), Compositional Data Analysis: Theory and Applications, Wiley, Chichester UK, (isbn 0-470-71135-3), ch. 2, 12-27.

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

[26] M. L. Eaton, 1983. Multivariate Statistics. A Vector Space Approach. John Wiley & Sons.

[27] P. Filzmoser, and K. Hron, 2008. Outlier detection for compositional data using robust methods. Mathematical Geosciences, 40, Number 3, 233-248.

[28] P. Filzmoser, and B. Steiger, 2009. StatDA: Statistical Analysis for Environmantel Data. R package version 1.1.

[29] R. Tolosana-Delgado, 2006. Geostatistics for constrained variables: positive data, compositions and probabilities. Applications to environmental hazard monitoring.PhD thesis, University of Girona, Girona (E).198 p.

[30] R. Tolosana-Delgado, V. Pawlowsky-Glahn, and J. J. Egozcue, 2008. Indicator kriging without order relation violations. Mathematical Geosciences, 40 Number 3, 327-347.

[31] R. Tolosana-Delgado, and H. von Eynatten, 2009. Grain-size control on petrographic composition of sediments: Compositional regression and rounded zeros. Mathematical Geosciences, 41, 869886.

[32] R. Gueorguieva, R. Rosenheck, and D. Zelterman, 2008. Dirichlet component regression and its applications to psychiatric data. Comput. Stat. Data Anal., 52, Number 12, 5344-5355.

[33] R. H. Hijazi, and R. W. Jernigan, 2009. Modelling compositional data using Dirichlet regression models. Journal of Applied Probability and Statistics. **in press**.

[34] S. Thió-Henestrosa, J. J. Egozcue, V. Pawlowsky-Glahn, L. O. Kovàcs, and G. P. Kovàcs, 2008. Balance-dendrogram. A new routine of CoDaPack. Computers and Geosciences, 34, 1682-1696.

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

[36] V. Pawlowsky-Glahn, and J. J. Egozcue, 2002. BLU estimators and compositional data. Mathematical Geology, 34 Number 3, 259-274.

[37] V. Pawlowsky-Glahn, and J. J. Egozcue, 2011. Exploring Compositional Data with the Coda-Dendrogram. Austrian Journal of Statistics, 40 Number (1-2), 103-113.

[38] V. Pawlowsky-Glahn, and R. A. Olea, 2004. Geostatistical Analysis of Compositional Data. Number 7, in Studies in Mathematical Geology. Oxford University Press.