MEASUREMENT SCIENCE REVIEW, **16**, (2016), No. 5, 235–243

DE GRUYTER OPEN

MEASUREMENT SCIENCE REVIEW

Journal homepage: http://www.degruyter.com/view/j/msr

ISSN 1335-8871

Practical Aspects of Log-ratio Coordinate Representations in Regression with Compositional Response

Eva Fišerová[1], Sandra Donevska[1], Karel Hron[1], Ondřej Bábek[2] and Kristýna Vaňkátová[1]

[1]*Department of Mathematical Analysis and Applications of Mathematics and* [2]*Department of Geology, Faculty of Science, Palacký University, 17. listopadu 12, 771 46 Olomouc, Czech Republic, eva.fiserova@upol.cz*

Regression analysis with compositional response, observations carrying relative information, is an appropriate tool for statistical modelling in many scientific areas (e.g. medicine, geochemistry, geology, economics). Even though this technique has been recently intensively studied, there are still some practical aspects that deserve to be further analysed. Here we discuss the issue related to the coordinate representation of compositional data. It is shown that linear relation between particular orthonormal coordinates and centred log-ratio coordinates can be utilized to simplify the computation concerning regression parameters estimation and hypothesis testing. To enhance interpretation of regression parameters, the orthogonal coordinates and their relation with orthonormal and centred log-ratio coordinates are presented. Further we discuss the quality of prediction in different coordinate system. It is shown that the mean squared error (MSE) for orthonormal coordinates is less or equal to the MSE for log-transformed data. Finally, an illustrative real-world example from geology is presented.

Keywords: Orthonormal coordinates, centred -ratio coordinates, orthogonal coordinates, log-transformation, multivariate regression model, mean squared error.

1. Introduction

Compositional data, observations carrying relative information, have been studied in the framework of the log-ratio methodology [1]. This approach enables to relax the fixed constant sum constraint of their proportional or percentage representations (1 and 100, respectively), and follow natural principles of compositions. They consist of scale invariance (results of a statistical analysis should be the same irrespective of concrete representation of the positive vector whose parts carry relative contributions on a whole), permutation invariance and subcompositional coherence [25]. Particularly, the issue of scale invariance is very important, because it enables to decide, just according to the nature of the studied problem, whether the relative structure of variables is the main focus of the analysis, or not. Even if not, relevant preprocessing of the original data is still a crucial point that should be taken into account [24]. Due to the algebraic-geometrical structure of compositional data [4, 23], called nowadays the Aitchison geometry [23] with Euclidean vector space properties, it is possible to express compositions in such real coordinates that enable to proceed with standard statistical processing without further constraints, by considering the specific interpretation of coordinates in sense of log-ratios (or, more general, in log-contrasts) of the original compositional parts. As usual in any reasonable statistical analysis, orthonormal coordinates are preferable [5], although also other coordinate representations are useful in some special cases.

All these aspects concern also regression with compositional response and real covariates, i.e., where not absolute values of response variables, but rather ratios between them form the source of primary information. The corresponding linear regression model has been extensively studied, both in terms of the original compositional response within the Aitchison geometry, and in any of the established log-ratio coordinate systems [8, 20, 25, 29, 30]. It turned out that working in orthonormal coordinates is a preferable option, although their particular interpretation in terms of balances between groups of compositional parts [7] needs to be taken into account. Regression analysis with compositional response is of great potential interest in geochemistry and also in medical applications, e.g., in human metabolomics, where concentrations of metabolites are frequently influenced by external factors (temperature, age of patients, etc.).

Despite this intensive care, there are still some practical aspects concerning linear regression with compositional response that deserve to be further analysed. The first one concerns special orthonormal coordinate systems that enable interpretation in terms of the original compositional parts (with respect to the other parts in the actual composition) and were applied in a number of applications including regression modelling [10, 12, 14]. Although it is theoretically sound to work exclusively in orthonormal coordinates, this particular choice

DOI: 10.1515/msr-2016-0029

of coordinates seems to be also a bit impractical as for a D-part composition D coordinate systems are needed. It can be shown that due to the relation between these particular orthonormal coordinates and centred log-ratio coordinates [1] that are formed by coefficients with respect to a generating system, it is possible to get easily the same numerical outputs (or possibly up to a constant multiple) concerning regression parameters estimation and hypotheses testing in multivariate regression models by using just one coordinate system. The second aspect concerns the relation between the mean square error (MSE) and the coefficient of determination, obtained from a regression model in orthonormal coordinates, or after applying a log-transformation to the original compositional data (in units that do not clearly indicate relative structure of components, like proportions or percentages). This fact may be useful for further methodological developments, similarly as it was the case of inequality between Euclidean distance in orthonormal log-ratio coordinates (or, equivalently, the Aitchison distance [2] between the original compositions) and Euclidean distance between log-transformed compositions [17]. For example, the mentioned relation was successfully used for the case of compositional data with an informative total (sum of parts), characterized by so called T-spaces [24], where a log-transformation plays an important role of a possible coordinate representation as an alternative to orthonormal log-ratio coordinates plus a variable representing the total.

The paper is organized as follows. In the next section some basics of the log-ratio methodology for compositional data, necessary for this contribution, are recalled. In Section 3 the same is done for multivariate linear regression models together with statistical inference on regression parameters that is not readily available in standard statistical literature. Detailed analysis of the possible use of clr coordinates for estimation and statistical inference in regression with compositional response follows in Section 4. Section 5 presents outputs on the MSE and the coefficient of determination for regression models in orthonormal log-ratio coordinates and for log-transformed compositions. An illustrative example from geochemistry is presented in Section 6 and the final Section 7 concludes.

2. Log-ratio approach to compositional data

In compositional data, the primary source of information is contained in (log-)ratios between components (parts). From a practical perspective it means that whenever not the absolute values of parts, but rather their relative structure is of interest, any particular unit representation of compositions (that can even vary in the data set) should not alter the results of their reasonable statistical processing; we refer to scale invariance of compositional data. From the perspective of geochemical data, any of their representations, either in mg/kg, ppm (since not all components need to be measured, i.e., no constant sum constraint is expected), proportions, or percentages, is fully equivalent. Together with other two more methodological principles, permutation invariance and subcompositional coherence [25], they form general requirements for any reasonable compositional data analysis. These principles are followed by the Aitchison geometry [23] that results for a D-part composition $\mathbf{z} = (z_1, \ldots, z_D)$ in a $(D-1)$-dimensional Euclidean vector space. As a consequence, it is possible to form coordinates with respect to an orthonormal basis, or generating system that map the Aitchison geometry to the standard Euclidean geometry in real space, for which most multivariate statistical methods are designed [5].

Centered log-ratio (clr) coordinates represent coefficients with respect to a generating system and historically the first isometric mapping between the Aitchison geometry and the real space endowed with the Euclidean geometry [1]. Clr coordinates are defined as

$$\mathrm{clr}(\mathbf{z}) = \left(\ln \frac{z_1}{\sqrt[D]{\prod_{j=1}^{D} z_j}}, \ldots, \ln \frac{z_D}{\sqrt[D]{\prod_{j=1}^{D} z_j}} \right). \qquad (1)$$

Clr coordinates are characterized by a zero sum of the variables and, consequently, by a singular covariance matrix. On the other hand, they are permutation invariant with respect to the original compositional parts; also an interpretation of clr coordinates in terms of dominance of single parts in the given composition is possible. Though, for the purpose of statistical processing it is preferable to have orthonormal coordinates that avoid singularity of the covariance matrix and guarantee isometry with the original sample space of compositions, the simplex (note that the isometry property holds also for the clr coordinates).

Geometrically, orthonormal log-ratio coordinates can be derived on a hyperplane, formed by clr coordinates. Any of its orthonormal bases can be expressed as rows of a $(D-1) \times D$ matrix $\mathbf{V}$ satisfying the relation $\mathbf{V}\mathbf{V}' = \mathbf{I}_{(D-1)}$. Here the symbol $'$ indicates transposition. Accordingly, orthonormal coordinates are obtained as

$$\mathbf{y} = \mathrm{clr}(\mathbf{z})\mathbf{V}', \qquad (2)$$

and any two orthonormal coordinate systems are just rotations of each other [6].

One popular choice of the orthonormal basis leads to the matrix $\mathbf{V} = \left(\mathbf{v}_1', \cdots, \mathbf{v}_{D-1}'\right)'$, where D-dimensional row vectors $\mathbf{v}_i$, $i = 1, \ldots, D-1$, are given by

$$\mathbf{v}_i = \sqrt{\frac{D-i}{D-i+1}} \left(0, \cdots, 0, 1, -\frac{1}{D-i}, \ldots, -\frac{1}{D-i} \right)$$

(the value 1 is placed in the ith position). Using this orthonormal basis leads to the following ilr coordinates

$$y_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{z_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} z_j}}, \; i = 1, \ldots, D-1. \qquad (3)$$

In this setting, the first coordinate y_1 explains all the relative information about the first compositional part z_1 within the given composition [12]. It can be interpreted in terms of dominance of a part in the numerator of the log-ratio with respect

to other parts in the composition, aggregated by their geometric mean. One should be just aware that y_1 can be never interpret as a coordinate carrying all the information contained in one component in an absolute sense; it is always referred to the composition or subcomposition considered. It also resembles the first clr coordinate, denoted as $\mathrm{clr}(\mathbf{z})_1$; indeed, the first coordinates of both systems are related through

$$\mathrm{clr}(\mathbf{z})_1 = \sqrt{\frac{D-1}{D}}\, y_1. \tag{4}$$

In order to obtain coordinates with similar interpretation for each of the compositional parts z_l, $l = 1,\ldots,D$, D different orthonormal coordinate systems are needed, where the D-part composition $(z_1,\ldots,z_D)$ in (3) is replaced by a permuted composition $\mathbf{z}^{(l)}$, $l = 1,2,\ldots,D$:

$$\begin{aligned}\mathbf{z}^{(l)} &= (z_l, z_1, \ldots, z_{l-1}, z_{l+1}, \ldots, z_D) = \\ &= (z_1^{(l)}, z_2^{(l)}, \ldots, z_{l-1}^{(l)}, z_l^{(l)}, z_{l+1}^{(l)}, \ldots, z_D^{(l)}).\end{aligned}$$

Accordingly, orthonormal coordinates $\mathbf{y}^{(l)}$ are obtained, where the first coordinate aggregates all log-ratios with the compositional part z_l [12]. Note that also the matrix relation between $\mathbf{y}$ and $\mathbf{y}^{(l)}$ is a result of a simple permutation operation. Thus, without loss of generality we can focus just on the case of (3) in the following.

3. Estimation in regression model with compositional response

Although it is possible to construct a regression model with compositional response directly for the original compositional data in the Aitchison geometry [8], any statistical inference needs to involve the whole compositional response. Instead, the regression model needs to be expressed in orthonormal coordinates, where standard testing procedures can be applied by considering the specific interpretation of balances [7] and their special cases [10].

Regression with a D-part compositional response leads to a multivariate linear model with a $(D-1)$-dimensional response variable. Although by using orthonormal coordinates, it is possible to decompose the multivariate model into $D-1$ multiple regressions [8], in general, the multivariate approach has several advantages in comparison with a series of univariate models. Multivariate models respect the association between outcomes, and thus, in general, procedures are more efficient. They can evaluate the joint influence of predictors on all outcomes and avoid the issue of multiple testing. Below we list some basics of multivariate linear regression models that are used in the following sections. As usual in the regression context, column vectors are considered here. A multivariate linear model [13] can be expressed in the form

$$(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{D-1}) = \mathbf{X}(\beta_1, \beta_2, \ldots, \beta_{D-1}) + (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_{D-1}),$$

or, equivalently, in the matrix form

$$\underline{\mathbf{Y}}_{(n\times(D-1))} = \mathbf{X}_{(n\times k)}\mathbf{B}_{(k\times(D-1))} + \underline{\varepsilon}_{(n\times(D-1))}.$$

The design matrix $\mathbf{X}$ is of full column rank, β_j, $j = 1,2,\ldots,D-1$, is a k-dimensional vector of unknown regression parameters and $\underline{\varepsilon}$ is a matrix of random errors. The multivariate responses $\underline{\mathbf{Y}}_{i\cdot} = (Y_{i1},\ldots,Y_{iD-1})'$, $i = 1,2,\ldots,n$, are independent with the same unknown variance-covariance matrix Σ, i.e.

$$\begin{aligned}\mathrm{cov}(\underline{\mathbf{Y}}_{i\cdot}, \underline{\mathbf{Y}}_{j\cdot}) &= \mathbf{0}_{((D-1)\times(D-1))},\ i \neq j,\\ \mathrm{var}(\underline{\mathbf{Y}}_{i\cdot}) &= \Sigma_{((D-1)\times(D-1))},\ i = 1,\ldots n.\end{aligned}$$

The best linear unbiased estimator (BLUE) of the parameter matrix $\mathbf{B}$

$$\widehat{\mathbf{B}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_{D-1})$$

is invariant under a change of the variance-covariance matrix Σ. However, the variance-covariance matrix of the vector $\mathrm{vec}(\widehat{\mathbf{B}}) = (\widehat{\beta}_1', \widehat{\beta}_2', \ldots, \widehat{\beta}_{D-1}')'$

$$\mathrm{var}\left[\mathrm{vec}(\widehat{\mathbf{B}})\right] = \Sigma \otimes \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

depends on Σ. Here the symbol $\otimes$ denotes the Kronecker product. Since the variance-covariance matrix Σ is unknown, it is necessary to estimate it. The unbiased estimator of Σ is $\widehat{\Sigma} = \underline{\mathbf{Y}}'\mathbf{M}_X\underline{\mathbf{Y}}/(n-k)$, where $\mathbf{M}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is a projector on the orthogonal complement of the vector space $\mathscr{M}(\mathbf{X})$ generated by the columns of the matrix $\mathbf{X}$, i.e. $\mathscr{M}(\mathbf{X}) = \{\mathbf{X}\mathbf{u} : \mathbf{u} \in \mathbb{R}^k\}$. Under normality, the estimators $\widehat{\mathbf{B}}$ and $\widehat{\Sigma}$ are statistically independent. Moreover, if $n-k > D-1$, then $(n-k)\widehat{\Sigma}$ has the Wishart distribution $W_{D-1}[n-k,\Sigma]$.

Let us note that the univariate approach leads to the same estimators of regression parameters β_j and of variances $\sigma_{jj} = \{\Sigma\}_{jj}$, $j = 1,2,\ldots,D-1$. Specifically, $\widehat{\beta}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_i$, $\mathrm{var}(\widehat{\beta}_i) = \sigma_{ii}(\mathbf{X}'\mathbf{X})^{-1}$, and $\widehat{\sigma}_{ii} = \mathbf{Y}_i'\mathbf{M}_X\mathbf{Y}_i/(n-k)$.

Under the assumption of normally distributed coordinate representation $\underline{\mathbf{Y}}_{i\cdot}$ of the compositional response [19], hypotheses testing can be performed. Theory of multivariate linear regression models [16] provide a range of tests, that are easy to compute due to explicit formulas and do not require to consider iterative algorithms under mild linearity assumptions.

Usually three basic cases of hypotheses testing in multivariate regression context are considered: significance of covariates for the ilr coordinate y_j, $j = 1,2,\ldots,D-1$, point wise and simultaneously, and verification that the regressor x_i, $i = 1,2,\ldots,k$, contributes to the explanation of the overall variability in $\underline{\mathbf{Y}}$.

It is easy to see that significance tests on single regression parameters as well as hypotheses testing on the whole vector parameter β_j, $j = 1,\ldots,D-1$, that conveys contributions of all covariates to the j-th coordinate simultaneously can be performed within univariate multiple regressions using standard $T-$ and $F-$ test statistics, respectively. Particularly, the test statistics for the null hypothesis $\beta_j = \mathbf{0}$ can be expressed as

$$F_j^{ilr} = \frac{(n-k)\,\widehat{\beta}_j'\mathbf{X}'\mathbf{X}\widehat{\beta}_j}{k\widehat{\sigma}_{jj}}, \tag{5}$$

which has F-distribution with k and $n-k$ degrees of freedom under the null hypothesis.

The case of significance testing of the i-th regressor, $i = 1,\ldots,k$, requires already the multivariate setting. Symbolically, the null hypothesis about the i-th regressor is expressed as $H_{0i} : \mathbf{B}_{i\cdot} = \left(\beta_{i1},\beta_{i2},\ldots,\beta_{i(D-1)}\right) = \mathbf{0}$. The corresponding test statistic is given by

$$F^{ilr}_{pred,i} = \frac{(n-D-k+2)\,\widehat{\mathbf{B}}_{i\cdot}\,(\underline{\mathbf{Y}}'\mathbf{M}_X\underline{\mathbf{Y}})^{-1}\,\widehat{\mathbf{B}}'_{i\cdot}}{(D-1)\left\{(\mathbf{X}'\mathbf{X})^{-1}\right\}_{ii}}, \qquad (6)$$

which is distributed as $F_{D-1,n-D-k+2}$ under the null hypothesis H_{0i}.

Finally, in some cases even significance of the whole matrix of regression parameters $\mathbf{B}$, or a more general hypothesis $H_0 : \mathbf{AB} = \mathbf{C}$, where $\mathbf{A}$ is a $q \times k$ hypothesis matrix having full-row rank $q \le k$, and $\mathbf{C}$ is a $q \times D-1$ matrix, are of interest as well. For this purpose a battery of tests is available in the literature [13], like Pillai-Bartlett trace, Wilks's Lambda, Hotelling-Lawley trace and Roy's largest root. All of them are based directly or indirectly on $p = \min(q, D-1)$ nonzero eigenvalues λ_j of the product matrix $\mathbf{HE}^{-1}$, where $\mathbf{H}$ is the matrix for the hypothesis sums of squares and cross products, and $\mathbf{E}$ is the residual matrix, i.e.

$$\begin{aligned}\mathbf{E} &= (\underline{\mathbf{Y}} - \mathbf{X}\widehat{\mathbf{B}})'(\underline{\mathbf{Y}} - \mathbf{X}\widehat{\mathbf{B}})\\ \mathbf{H} &= (\mathbf{A}\widehat{\mathbf{B}} - \mathbf{C})'[\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}']^{-1}(\mathbf{A}\widehat{\mathbf{B}} - \mathbf{C}).\end{aligned}$$

The behaviour of these matrices in different coordinate systems is thus crucial for statistical properties of the above test statistics. Obviously, all of them are invariant under a change of a basis thus they follow the behaviour of the sample covariance matrix under affine transformations [18].

4. Regression with compositional response in different coordinate systems

Due to (4) that describes the relationship between single clr coefficients and the first ilr coordinates from (3) it seems to be quite intuitive possibility to replace orthonormal coordinates in the response simply by their clr counterparts and then proceed with the regression analysis.

Nevertheless, due to singularity of the covariance matrix of clr coordinates it is not possible to decompose the multivariate model into univariate ones as it was the case for orthonormal coordinates. Though, as it is shown below, even taking multivariate regression in clr coordinates would yield the same results of the respective test statistics as one would obtain by considering single ilr coordinates, coming from D regression models.

In the following, the relation between clr and ilr coordinate systems (2) is extensively used. Denote clr coordinates of the composition $\mathbf{z}$ as $\mathbf{y}^{clr}$. Then the multivariate model can be also written in the form

$$\underline{\mathbf{Y}}^{clr} = \mathbf{XB}^{clr} + \underline{\varepsilon}_{clr}. \qquad (7)$$

The variance-covariance matrix of independent D-variate responses $\underline{\mathbf{Y}}^{clr}_{i\cdot}$ is $\mathrm{var}(\underline{\mathbf{Y}}^{clr}_{i\cdot}) = \Sigma_{clr} = \mathbf{V}'\Sigma_{ilr}\mathbf{V}$, $i = 1,\ldots n$. The variance-covariance matrix Σ_{clr} is a $D \times D$ positive semi-definite matrix with the rank $D-1$ unlike Σ_{ilr}, which is a full rank $(D-1)\times(D-1)$ positive definite matrix. Obviously, $\Sigma_{ilr} = \mathbf{V}\Sigma_{clr}\mathbf{V}'$. The relationships between the parameter matrices and multivariate responses are the following

$$\begin{aligned}\mathbf{B}^{clr} &= \mathbf{B}^{ilr}\mathbf{V},\\ \mathbf{B}^{ilr} &= \mathbf{B}^{clr}\mathbf{V}',\\ \underline{\mathbf{Y}}^{clr} &= \underline{\mathbf{Y}}^{ilr}\mathbf{V},\\ \underline{\mathbf{Y}}^{ilr} &= \underline{\mathbf{Y}}^{clr}\mathbf{V}'.\end{aligned} \qquad (8)$$

Theorem 1. *(i) The test statistics for the hypotheses $\mathbf{B}^{ilr}_{i\cdot} = \mathbf{0}$ and $\mathbf{B}^{clr}_{i\cdot} = \mathbf{0}$ are the same for an arbitrary $i = 1,2,\ldots,k$, i.e. $F^{ilr}_{pred,i} = F^{clr}_{pred,i}$.*

(ii) Let us denote $\beta_{clr,l}$ the l-th column vector of the parameter matrix $\mathbf{B}^{clr}$ in the model with clr coordinates responses, and $\beta^{(l)}_{ilr,1}$ the first column vector of the parameter matrix $\mathbf{B}^{ilr}$ in the l-th model with orthonormal coordinates $\mathbf{y}^{(l)}$ considered as multivariate responses. Then the test statistics for the null hypotheses $\beta^{(l)}_{ilr,1} = \mathbf{0}$ and $\beta_{clr,l} = \mathbf{0}$ for an arbitrary $l = 1,2,\ldots,D$, are the same, i.e. $F^{ilr,(l)}_1 = F^{clr}_l$.

Proof. Let us consider the first statement. According to the relations (4) and (8), as well as from the fact that the matrix $\widehat{\Sigma}_{clr}$ is singular with the rank $D-1$, the test statistic $F^{clr}_{pred,i}$ that arises from a general formula in [16] can be rewritten as

$$\begin{aligned}F^{clr}_{pred,i} &= \frac{(n-\mathrm{r}(\mathbf{X})-\mathrm{r}(\widehat{\Sigma}_{clr})+1)\widehat{\mathbf{B}}^{clr}_{i\cdot}\widehat{\Sigma}^{-}_{clr}(\widehat{\mathbf{B}}^{clr}_{i\cdot})'}{\mathrm{r}(\widehat{\Sigma}_{clr})\{(\mathbf{X}'\mathbf{X})^{-1}\}_{ii}}\\ &= \frac{(n-D-k+2)\widehat{\mathbf{B}}^{ilr}_{i\cdot}\mathbf{V}(\mathbf{V}'\widehat{\Sigma}_{ilr}\mathbf{V})^{-}\mathbf{V}'(\widehat{B}^{ilr}_{i\cdot})'}{(D-1)\{(\mathbf{X}'\mathbf{X})^{-1}\}_{ii}}\\ &= \frac{(n-D-k+2)\widehat{\mathbf{B}}^{ilr}_{i\cdot}\mathbf{V}\mathbf{V}^{-}_{L}\widehat{\Sigma}^{-1}_{ilr}(\mathbf{V}'_R)^{-}\mathbf{V}'(\widehat{\mathbf{B}}^{ilr}_{i\cdot})'}{(D-1)\{(\mathbf{X}'\mathbf{X})^{-1}\}_{ii}},\end{aligned}$$

where the matrix $\mathbf{V}'_R$ is the right inverse of $\mathbf{V}'$ and the matrix $\mathbf{V}_L$ is the left inverse of $\mathbf{V}$, i.e.,

$$\left(\mathbf{V}'_R\right)^{-} = \widehat{\Sigma}_{ilr}\mathbf{V}\left(\mathbf{V}'\widehat{\Sigma}_{ilr}\mathbf{V}\right)^{-} \quad \text{and} \quad \mathbf{V}'\left(\mathbf{V}'_R\right)^{-} = \mathbf{I}_D,$$

$$\mathbf{V}^{-}_{L} = \left(\mathbf{V}'\widehat{\Sigma}_{ilr}\mathbf{V}\right)^{-}\mathbf{V}'\widehat{\Sigma}_{ilr} \quad \text{and} \quad \mathbf{V}^{-}_{L}\mathbf{V}' = \mathbf{I}_{D-1}$$

and $\mathbf{A}^{-}$ denotes a generalized inverse of a matrix $\mathbf{A}$, i.e., a matrix fulfilling the property $\mathbf{AA}^{-}\mathbf{A} = \mathbf{A}$.

The desired equality $F^{ilr}_{pred,i} = F^{clr}_{pred,i}$ is gained by pre-multiplying and post-multiplying the matrix $\widehat{\Sigma}_{ilr}$ by $\mathbf{VV}' = \mathbf{I}_{D-1}$.

The statement (ii) is a direct consequence of (4). □

Theorem 2. *The test statistics for the null hypotheses $\mathbf{B}^{ilr} = \mathbf{0}$ and $\mathbf{B}^{clr} = \mathbf{0}$, as listed in Section 3, are the same.*

Proof. The statement follows from invariance under a change of a basis of the matrices $\mathbf{E}$ and $\mathbf{H}$, $\mathbf{E}_{ilr} = \mathbf{VE}_{clr}\mathbf{V}'$, $\mathbf{E}_{clr} = \mathbf{V}'\mathbf{E}_{ilr}\mathbf{V}$, $\mathbf{H}_{ilr} = \mathbf{VH}_{clr}\mathbf{V}'$, $\mathbf{H}_{clr} = \mathbf{V}'\mathbf{H}_{ilr}\mathbf{V}$, and the fact that the matrices $\mathbf{H}_{clr}\mathbf{E}^{-}_{clr}$ and $\mathbf{H}_{ilr}\mathbf{E}^{-1}_{ilr}$ have the same non-zero eigenvalues. □

The above findings can be used to perform parameter estimation and significance testing in clr coordinates instead of taking D orthonormal coordinate systems of type (3), when the interpretation in sense of the original compositional parts (with respect to the others) is required. Although methodically working in orthonormal coordinates is preferred in any case, numerical outputs are the same (test statistics) or differ just up to a constant resulting from (4).

Finally, note that the interpretation of the regression parameters can be enhanced by considering orthogonal coordinates, resulting from suppressing scaling constants in orthonormal coordinates. Concretely, they are formed from (3) by omitting scaling constants and replacing the natural logarithm by its binary counterpart (or any other interpretable base of logarithm), i.e.

$$y_i^* = \log_2 \frac{z_i}{\sqrt[D-i]{\prod_{j=i+1}^{D} z_j}},\ i = 1,\ldots,D-1 \qquad (9)$$

[22]. By considering regression in clr coordinates, the parameters of the resulting regression model in orthogonal coordinates, adapted to favour the l-th compositional part (denoted as $\beta_1^{*(l)}$), would be related through

$$\beta_1^{*(l)} = \log_2(\mathrm{e})\sqrt{\frac{D}{D-1}}\beta_{ilr,1}^{(l)} = \log_2(\mathrm{e})\frac{D}{D-1}\beta_{clr,l}. \qquad (10)$$

Consequently, by taking the j-th element of $\beta_1^{*(l)}$, i.e. $\beta_{1;j}^{*(l)}$, for $j = 1,\ldots,k$, then for a unit additive change in the j-th explanatory variable (by constant values of the other covariates), the ratio of x_l to the mean relative contributions of the other parts grows (decreases) $\delta = 2^{\beta_{1;j}^{*(l)}}$ times.

5. Quality of prediction in log-ratio coordinates versus log-transformed data

In practice, the simple log-transformation, $y_i = \log(z_i)$, $i = 1,2,\ldots,D$, is often used in geochemistry, chemometrics and related fields for modelling data with strictly positive parts. Nevertheless, it has important consequences also in the compositional context. If not just the relative structure of compositional parts is of interest, but also their absolute abundances in the original units like mg/l, cps, or monetary units [24], the log-transformation serves for an appropriate coordinate representation of the data at hand. Namely, compositional data with informative absolute values of parts induce an Euclidean vector space structure again (we refer to T-space) that should be taken into account for the construction of any relevant real coordinates.

An obvious consequence in the case of positive data (i.e., compositional data with an informative total) is lack of scale invariance, but relative scale of compositions (not absolute differences, but ratios form the source of dissimilarity between compositional vectors) is still taken into account for statistical processing. Interestingly, it is easy to see that the standard Euclidean distance of log-transformed data is always greater or equal to the Aitchison distance between two compositions $\mathbf{x}$ and $\mathbf{y}$ [2], defined as

$$d_a(\mathbf{x},\mathbf{y}) = \sqrt{\frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\ln\frac{x_i}{x_j} - \ln\frac{y_i}{y_j}\right)^2}.$$

To compare log-ratio and log-transformed regression models one has to analyse, whether something similar holds also in the regression context. Such a finding would be an important step to understand the behaviour of regression models in different coordinate systems. For this purpose, the matrix of sums of residual squares is taken for both the cases of ilr coordinates and log-transformed compositions,

$$\begin{aligned}\mathbf{E}_{ilr} &= (\underline{\mathbf{Y}} - \mathbf{X}\widehat{\mathbf{B}})'(\underline{\mathbf{Y}} - \mathbf{X}\widehat{\mathbf{B}}) = \underline{\mathbf{Y}}'\mathbf{M}_X\underline{\mathbf{Y}},\\ \mathbf{E}_{log} &= [\log(\underline{\mathbf{Z}})]'\mathbf{M}_X\log(\underline{\mathbf{Z}}),\end{aligned}$$

respectively. Here the symbol $\underline{\mathbf{Z}}$ denotes an $n \times D$ matrix with D-part compositions in rows. The overall variability in data corresponds to the matrices of total sum of squares

$$\mathbf{T}_{ilr} = \underline{\mathbf{Y}}'\mathbf{M}_1\underline{\mathbf{Y}} = \mathbf{V}\mathbf{T}_{log}\mathbf{V}', \quad \mathbf{T}_{log} = [\log(\underline{\mathbf{Z}})]'\mathbf{M}_1\log(\underline{\mathbf{Z}}).$$

The matrix $\mathbf{E}$ is commonly used to measure the discrepancy between the data and a fitted model in case of multivariate regression [13]. Although also an alternative exists, based directly on the norm between the observed and predicted response [8, 30], using directly $\mathbf{E}$ seems to be more coherent with the current regression methodology. Particularly, the trace of $\mathbf{E}$ is of primary importance, because it aggregates residual sums of squares of single response variables and leads to the multivariate analogy of the residual sum of squares (RSS). The inequalities between the traces of matrices $\mathbf{E}$ and $\mathbf{T}$ for compositions in ilr coordinates and by taking log-transformation are stated in the following theorem.

Theorem 3. *The traces of the matrices $\mathbf{E}_{ilr}$ (sums of residual squares) and $\mathbf{T}_{ilr}$ (total sum of squares) for compositions represented in ilr coordinates are always less or equal than the traces of the matrices $\mathbf{E}_{log}$ and $\mathbf{T}_{log}$ for log-transformed compositions, i.e. the following inequalities hold*

$$0 \le \mathrm{tr}(\mathbf{E}_{ilr}) \le \mathrm{tr}(\mathbf{E}_{log}), \quad 0 \le \mathrm{tr}(\mathbf{T}_{ilr}) \le \mathrm{tr}(\mathbf{T}_{log}). \qquad (11)$$

Proof. The relationships between the ilr, clr coordinates and log-transformations [1, 6] can be expressed as

$$\underline{\mathbf{Y}} = \mathrm{clr}(\underline{\mathbf{Z}})\mathbf{V}', \quad \mathrm{clr}(\underline{\mathbf{Z}}) = \mathbf{M}_1\log(\underline{\mathbf{Z}}),$$

where $\mathbf{V}$ contains in its rows orthonormal basis in clr coordinates, i.e. it is a $(D-1)\times D$ matrix with the property $\mathbf{V}\mathbf{V}' = \mathbf{I}_{D-1}$, and $\mathbf{M}_1$ is a projection matrix on the orthogonal complement of the vector space $\mathscr{M}(\mathbf{1}) \subset \mathbb{R}^D$ generated by the vector $\mathbf{1}$ of n ones, i.e., on the hyperplane formed by clr coordinates. Using these equalities, the matrix $\mathbf{E}_{ilr}$ can be rewritten as

$$\begin{aligned}\mathbf{E}_{ilr} &= \mathbf{V}[\mathrm{clr}(\underline{\mathbf{Z}})]'\mathbf{M}_X\mathrm{clr}(\underline{\mathbf{Z}})\mathbf{V}'\\ &= \mathbf{V}\mathbf{M}_1[\log(\underline{\mathbf{Z}})]'\mathbf{M}_X\log(\underline{\mathbf{Z}})\mathbf{M}_1\mathbf{V}'.\end{aligned}$$

The matrix $\mathbf{V}$ contains basis of the vector space that is orthogonal to the vector space $\mathscr{M}(\mathbf{1})$, and thus $\mathbf{M}_1\mathbf{V}' = \mathbf{V}'$. Hence

$$\mathbf{E}_{ilr} = \mathbf{V}[\log(\underline{\mathbf{Z}})]'\mathbf{M}_X \log(\underline{\mathbf{Z}})\mathbf{V}' = \mathbf{V}\mathbf{E}_{log}\mathbf{V}',$$

and the trace of the matrix $\mathbf{E}_{ilr}$ is $\mathrm{tr}(\mathbf{E}_{ilr}) = \mathrm{tr}(\mathbf{E}_{log}\mathbf{V}'\mathbf{V})$. The matrix $\mathbf{E}_{log}$ is positive semidefinite, $\mathbf{V}'\mathbf{V}$ is symmetric, and thus, the upper and lower bounds for $\mathrm{tr}(\mathbf{E}_{ilr})$ are [15]

$$\lambda_{\min}(\mathbf{V}'\mathbf{V})\mathrm{tr}(\mathbf{E}_{log}) \le \mathrm{tr}(\mathbf{E}_{ilr}) \le \lambda_{\max}(\mathbf{V}'\mathbf{V})\mathrm{tr}(\mathbf{E}_{log}),$$

where $\lambda_{\min}$ and $\lambda_{\max}$ are the smallest and largest eigenvalues of the matrix $\mathbf{V}'\mathbf{V}$. Since the matrix $\mathbf{V}'\mathbf{V}$ is idempotent with the rank $D-1$, it has $D-2$ eigenvalues $\lambda_{max} = 1$ and one eigenvalue $\lambda_{\min} = 0$ [11]. Thus we have

$$0 \le \mathrm{tr}(\mathbf{E}_{ilr}) \le \mathrm{tr}(\mathbf{E}_{log}).$$

Similarly we can prove the inequality for the trace of matrices of total sum of squares. □

Theorem 3 states that the trace of the matrix $\mathbf{E}$ obtained for ilr coordinates is less or equal to that one for log-transformed compositions. Thus, the mean squared error (MSE) for ilr coordinates is less or equal to the MSE for log-transformed data. Since the same inequality holds also for the trace of the matrix $\mathbf{T}$, the relationship between the coefficients of determination R^2_{ilr} and R^2_{log} does not exist in general. These measures of goodness of fit, defined as

$$R^2_{ilr} = 1 - \frac{\mathrm{tr}(\mathbf{E}_{ilr})}{\mathrm{tr}(\mathbf{T}_{ilr})}, \quad R^2_{log} = 1 - \frac{\mathrm{tr}(\mathbf{E}_{log})}{\mathrm{tr}(\mathbf{T}_{log})},$$

thus reflect structural changes that arise by avoiding the scale invariance property of compositions, i.e. when log-transformation is applied instead of taking the ilr coordinates.

It is not difficult to demonstrate that there is no relation in general between both coefficients R^2_{ilr} and R^2_{log}. For this purpose, let us consider two matrices of response compositions,

$$\underline{\mathbf{Z}}_1 = \begin{pmatrix} 1 & 5 & 1 \\ 9 & 2 & 2 \\ 1 & 8 & 3 \\ \mathbf{1} & 2 & 5 \end{pmatrix}, \quad \underline{\mathbf{Z}}_2 = \begin{pmatrix} 1 & 5 & 1 \\ 9 & 2 & 2 \\ 1 & 8 & 3 \\ \mathbf{10} & 2 & 5 \end{pmatrix},$$

observed for the values $x = 1,2,3,4$ of the explanatory variable. Note that both matrices differ just by the entry on the position (4,1). Though, by taking linear regression with an absolute term, the first case results in $R^2_{ilr} = 0.707 < 0.736 = R^2_{log}$, while in the second one $R^2_{ilr} = 0.788 > 0.674 = R^2_{log}$ is obtained.

Finally, it is worth to mention that the trace of any covariance matrix (residual or total) is equal to the mean of the distances between the samples and the centre on the simplex. This fact can be used to reformulate Theorem 3 in terms of distances in the respective spaces, if appropriate.

6. Illustrative example: Reservoir Sediments in the Czech Republic

The findings from the above sections are briefly illustrated using a geological data set from lacustrine sediments of the Nové Mlýny reservoir in the Czech Republic (underwater core NM1, WGS-84: 48°53′8.771″N, 16°31′52.966″E) [26].

Thirty-four samples from the core were air dried, manually ground in agate mortar and subjected to element composition analysis using Energy Dispersive X-ray Fluorescence (EDXRF) spectrometry. A PANalytical MiniPal 4.0 EDXRF spectrometer with a Peltier-cooled silicon drift energy dispersive detector (Institute of Anorganic chemistry in Řež, Prague) was used. Signals of Al and Si were acquired at 4 kV/200 μA with Kapton filter 151 under He flush; Zn, Mn and Fe at 20 kV/100 μA with Al filter in air 152 and Rb and Pb at 30 kV/200 μA with Ag filter in air [21]. The EDXRF results are provided in counts per second (cps).

Fifteen elements Al, Si, P, Ti, K, Ca, Fe, Cr, Mn, Ni, Cu, Zn, Zr, Rb and Pb were selected for further statistical processing using regression analysis. The elements represent common lithophile elements, which are used for geochemical description of common parameters of sediments and sedimentary rocks, such as the grain size (Al, Si and Ti), degree of weathering (K, Al and Rb), heavy-mineral composition (Zr, Ti, Fe), organic production (P, Ca, Cu, Zn), redox state (Mn, Ni, Cu, Zn) and anthropogenic impact by toxic compounds (Cr, Ni, Zn, Pb).

In this concrete case, both absolute and relative information were of simultaneous interest; the total concentrations of the elements in the Nové Mlýny reservoir have been recently interpreted in [3]. Accordingly, in the following both log-ratio coordinates and log-transformed compositions were employed.

In addition to other site-specific geological tasks the aim was to investigate whether the distribution of these 15 elements in the core is random or organized. For this purpose linear regression models with the polynomial trend (up to the 4th-degree) in depth, and with the response composition in clr coordinates and log-transformed variables were taken. Particularly, the models (7) and

$$\log(\underline{\mathbf{Z}}) = \mathbf{X}\mathbf{B}^{log} + \underline{\varepsilon}_{log},$$

where $\mathbf{B}^{log} = (\beta_{log,1}, \ldots, \beta_{log,15})$, were analysed. The j-th row of the design matrix $\mathbf{X}$ was considered in the following forms

$$(1, \mathrm{depth}_j), \ \ldots, (1, \mathrm{depth}_j, \ldots, \mathrm{depth}_j^4).$$

In all cases the simplest possible model that was consistent with data was chosen.

By considering the regression outputs (realizations of test statistics F_l^{clr} and F-statistics to verify significance of the whole vector parameter $\beta_{clr,l}$ and $\beta_{log,l}$, respectively, T-statistics for significance testing of single regression parameters, p-values, coefficients of determination and visualization of data together with the corresponding regression functions), only zirconium (Zr) didn't show any systematic pattern (i.e. does not change with changing depth) either for log-transformation or clr coordinate of the response. A systematic increase/decrease was observed in a majority of the elements but their clr coordinates usually indicate a more complex (polynomial) underlying pattern.

A typical example is Fe (Figures 1, 4) in which an increasing trend was observed. For an easier interpretation, the response was expressed in orthogonal coordinates. From regression outputs (the slope parameter estimate was $8.796 \cdot 10^{-3}$ with standard error $0.710 \cdot 10^{-3}$ and p-value $\ll 0.001$, $MSE_{ilr} = 0.049$, $R^2_{ilr} = 0.845$) it can be concluded that by the increase of depth by 1 cm the ratio of Fe to the geometric mean of the other 14 elements increases approximately once ($\delta = 1.009146$ times, it means 1%); similarly, by considering log-transformed response (the slope parameter estimate was $3.766 \cdot 10^{-3}$ with standard error $0.580 \cdot 10^{-3}$ and p-value $\ll 0.001$, $MSE_{log} = 0.060$, $R^2_{log} = 0.609$), the increase of depth by 1 cm means that the absolute amount of Fe (in cps) grows approximately once, $\exp\{3.766 \cdot 10^{-3}\} = 1.003773$. From the lower value of MSE_{ilr} than MSE_{log} as indicated by Theorem 3 one can in general conclude that for given scales the *MSE* values show always better fit in the ilr space (that would be no longer the case for scaling-free R^2 values). On the other hand, data configuration for the clr representation suggests that the linear trend could be enhanced by a more complex regression function, here polynomial of degree four. An extreme case of this general feature is Si (Figures 2, 5), in which the linear trend for the response in clr coordinates is replaced by the polynomial one of degree four.

It is important to mention that the depth range from 45 to 55 cm in the NM1 core is a transitional zone between lower pre-dam fluvial sediments and upper, fully dam-reservoir ones [27]. This layer, strongly enriched in organic carbon, has critical effect on the depth distribution of various elements, including P (sensitive to organic productivity), Si (sensitive to grain size) and Fe (sensitive to redox conditions) (Figures 1-6). Distribution of these elements shows breaks at the base or on top of this layer, i.e. at 55 or 45 cm depth, which can be explained by their different geochemical behaviour. In particular, Si break is related to decrease of grain size at a break in sedimentation style from fluvial to lacustrine, Fe peak between 45 and 55 cm depth is likely related to diagenetic sulphide precipitation under dysoxic/anoxic conditions (high organic carbon) and P is related to increased organic productivity in water column of the lake.

Consequently, linear (= continuous in depth) regression trends are less likely than those represented by a polynomial function (= discontinuous in depth). In this respect, the clr data provide a better representation of the core stratigraphy. Mathematically, this effect can be easily explained by the remaining elements in the composition, which are incorporated in the denominator of the centred log-ratio. This facilitates identifying geochemical patterns related to the geochemical matrix in which the particular element is contained. On the other hand, there are also some exceptions, like for phosphorus (P, Figures 3, 6), where this change seems to be better reflected by the log-transformed response (accordingly, even two constant lines instead of one regression line were taken). In this case, the piecewise constant model with the j-th row of the design matrix given as $(1, I[\text{depth}_j \geq 45 \text{ cm}])$, where the symbol $I[\text{depth}_j \geq 45 \text{ cm}]$ denotes a dummy variable coded 1 for the j-th measurements in the depth 45 cm and more, and 0 otherwise, fitted best the data.

Based on the purpose of the analysis, one can consider purely relative information, or to take also absolute abundances of positive data into account. Nevertheless, like here, such decision of the analyst should always follow also previous expert knowledge on possible underlying processes in data.

7. Conclusions

Although regression analysis with compositional response represents one of the most tasks of compositional data analysis, there are still some aspects that deserve to be analysed in more detail. Two of them, concerning

- the particular coordinate representation for estimation and interpretation of regression parameters,
- the quality of prediction by considering (or not) also absolute abundances instead of purely relative information conveyed by compositional data

were discussed in this paper. They both have in common that even coordinate systems that are nowadays rather suppressed in compositional data analysis, here clr coordinates and log-transformed variables, might be useful for some specific tasks and also help to understand differences between various methodological viewpoints. Particularly, clr coordinates simplify the computation of the regression coefficients instead of considering D ilr regression models, just the principal difference between both options arising from a singularity of a covariance matrix for clr coordinates needs to be taken into account. Clr coordinates cannot be considered separately due to their zero sum constraint, while this is not the case for orthonormal (orthogonal) coordinates. The theoretical part of the paper was endowed with a real data example from sedimentology, where interesting patterns were revealed. From this perspective, we believe that the presented methodological outputs are useful steps for a practical analysis of compositional data.

Acknowledgment

The authors are grateful to the referees for helpful comments and suggestions. The authors gratefully acknowledge the support by the Operational Program Education for Competitiveness - European Social Fund (project CZ.1.07/2.3.00/20.0170 of the Ministry of Education, Youth and Sports of the Czech Republic), the grant COST Action CRoNoS IC1408, and the Grant IGA_PrF_2016_025 of the Internal Grant Agency of the Palacký University in Olomouc.

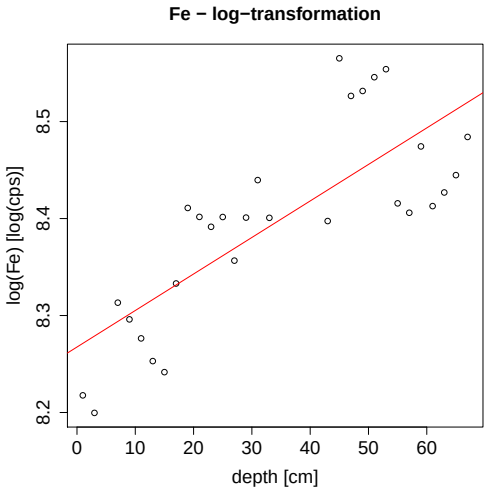

Fig. 1: Regression model for iron using log-transformation.

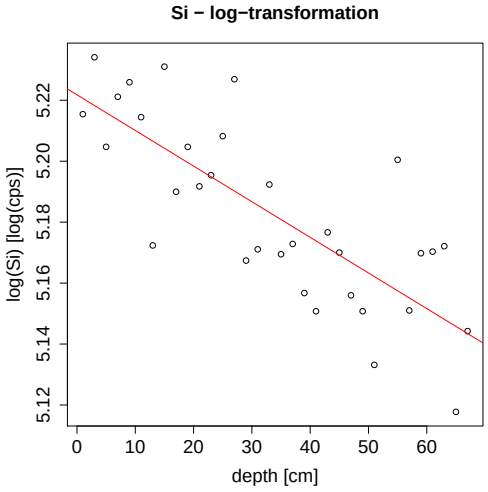

Fig. 2: Regression model for silicon using log-transformation.

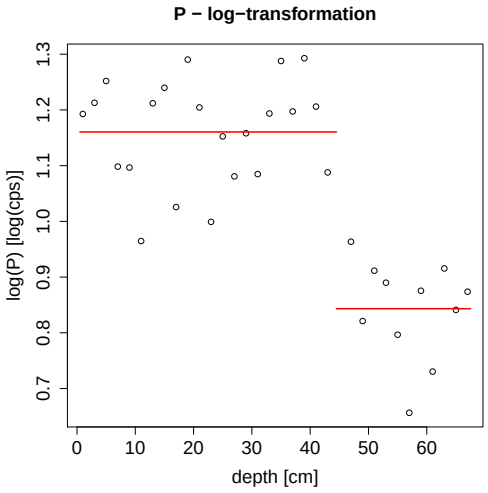

Fig. 3: Regression model for phosphorus using log-transformation.

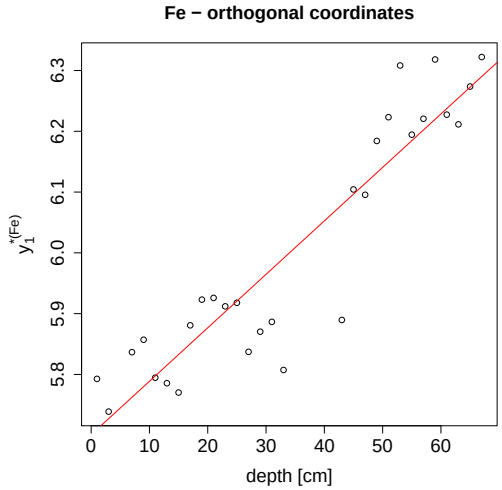

Fig. 4: Regression model for iron using orthogonal coordinates.

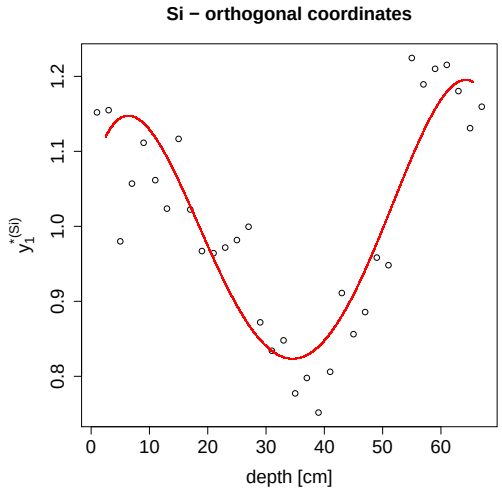

Fig. 5: Regression model for silicon using orthogonal coordinates.

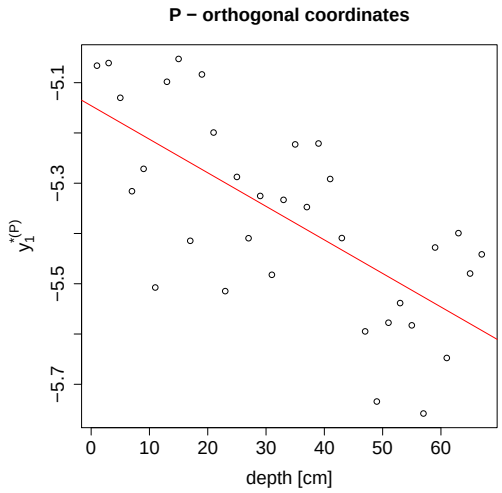

Fig. 6: Regression model for phosphorus using orthogonal coordinates.

REFERENCES

[1] Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall (Reprinted in 2003 with additional material by The Blackburn Press).

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

[3] Bábek, O., Matys Grygar, T., Faměra, M., Hron, K., Nováková, T., Sedláček, J. (2015). Geochemical background in polluted river sediments: How to separate the effects of sediment provenance and grain size with statistical rigour? *Catena*, 135, 240–253.

[4] Billheimer, D., Guttorp, P., Fagan, W. (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association*, 96(456), 1205–1214.

[5] Eaton, M.L. (1983). *Multivariate Statistics: A Vector Space Approach*. John Wiley & Sons.

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

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

[8] Egozcue, J.J., Pawlowsky-Glahn, V., Daunis-i-Estadella, J., Hron, K., Filzmoser, P. (2012). Simplicial regression. The normal model. *Journal of Applied Probability and Statistics*, 6, 87–106.

[9] Filzmoser, P., Hron, K. (2015). Robust coordinate for compositional data using weighted balances. In *Modern nonparametric, robust and multivariate methods*. Springer, 167–184.

[10] Ferrer-Rosell, B., Coenders, G., Mateu-Figueras, G., Pawlowsky-Glahn, V. (2016). Understanding low cost airline users' expenditure pattern and volume. *Tourism Economics*, 22, 269–291.

[11] Harville, D.A. (1997). *Matrix Algebra From a Statistician's Perspective*. Springer.

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

[13] Johnson, R.A, Wichern, D.W (2007). *Applied Multivariate Statistical Analysis* (6th Edition). Pearson.

[14] Kalivodová, A., Hron, K., Filzmoser, P., Najdekr, L., Janečková, H., Adam, T. (2015). PLS-DA for compositional data with application to metabolomics. *Journal of Chemometrics*, 29(1), 21–28.

[15] Kleinman, D.L., Athans, M. (1968). The design of suboptimal linear time-varying systems. *IEEE Transactions on Automatic Control*, AC-13, 150–159.

[16] Kubáček, L. (2008). *Multivariate statistical models revisited*. Olomouc, Czech Republic: Palacký University.

[17] Lovell, D., Müller, W., Taylor, J., Zwart, A., Helliwell, C. (2011). Proportions, percentages, PPM: Do the molecular biosciences treat compositional data right? In *Compositional data analysis: Theory and applications*. Wiley, 193–207.

[18] Martín-Fernández, J.A., Daunis-i-Estadella, J., Mateu-Figueras, G. (2015). On the interpretation of differences between groups for compositional data. *Statistics and Operations Research Transactions*, 39, 231–252.

[19] Mateu-Figueras, G., Pawlowsky-Glahn, V. (2008). Critical approach to probability laws in geochemistry. *Mathematical Geosciences*, 40(5), 489–502.

[20] Mateu-Figueras, G., Pawlowsky-Glahn, V., Egozcue, J.J. (2011). The principle of working on coordinates. In *Compositional data analysis: Theory and applications*. Wiley, 31–42.

[21] Matys Grygar, T., Elznicová, J., Bábek, O., Hošek, M., Engel, Z., Kiss, T. (2014). Obtaining isochrones from pollution signals in a fluvial sediment record: A case study in a uranium-polluted floodplain of the Ploučnice River, Czech Republic. *Appl Geochem*, 48, 1–15.

[22] Müller, I., Hron, K., Fišerová, E., Šmahaj, J., Cakirpaloglu, P., Vančáková, J. (2016). Time budget analysis using logratio methods. *arXiv:1609.07887* [math.ST].

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

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

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

[26] Sedláček, J., Bábek, O., Kielar, O. (2016). Sediment accumulation rates and high-resolution stratigraphy of recent fluvial suspension deposits in various fluvial settings, Morava River catchment area, Czech Republic. *Geomorphology*, 254, 73–87.

[27] Sedláček, J., Bábek, O., Nováková, T. (2016). Sedimentary record and anthropogenic pollution of a step-wise filled, multiple source fed dam reservoir: An example from Nové Mlýny reservoir, Czech Republic. *Science of the Total Environment*, DOI: 10.1016/j.scitotenv.2016.08.127.

[28] Templ, M., Hron, K., Filzmoser, P. (2016). Exploratory tools for outlier detection in compositional data with structural zeros. *Journal of Applied Statistics. DOI: 10.1080/02664763.2016.1182135*.

[29] van den Boogaart, K.G., Tolosana-Delgado, R. (2013). *Analyzing Compositional Data with R*. Springer.

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

Received June 1, 2016.
Accepted October 17, 2016.