Estimation of values above an upper detection limit in compositional data

D. Mikšová and P. Filzmoser

Vienna University of Technology

dominika.miksova@tuwien.ac.at

Motivation

- Geochemical compositions (e.g. concentrations of chemical elements in plants, soil and snow) are often affected by values **exceeding an upper detection limit** (UDL), besides that also by **rounded zeros** - values below lower detection limit.
- UDLs are common for K, P, Mn and Zn in ashed concentrations.
- For Compositional Data (CoDa), only the ratios between the variables (parts) contain the relevant information.
- "Advanced" imputation techniques make use of the multivariate information: CoDa deals with a specific geometry in the simplex, i.e. the **Aitchison geometry** preserving all metric properties.
- Imputation is necessary for statistical analyses that rely on complete data. Therefore, the task is to replace these values by meaningful numbers corresponding to the multivariate data structure.
- Simple (but naive) approach commonly used in practise: Replacing values above UDL by **1.2 times** the UDL.

CoDa transformation and Tobit regression

Truncated regression model (τ is truncation point):

$$E[y \mid y > \tau] = \mathbf{x}^t \cdot \beta + \sigma \left[\frac{\phi(\frac{\tau - \mathbf{x}^t \cdot \beta}{\sigma})}{1 - \Phi(\frac{\tau - \mathbf{x}^t \cdot \beta}{\sigma})} \right], \quad (1)$$

where ϕ and Φ are density and distribution function of $N(0,1)$, respectively.
$\rightarrow$ A complete version of the data matrix is required as a starting point for the iterative algorithm, therefore the initialization values above UDL is achieved by naive imputation.
$\rightarrow$ **x** is a composition, therefore it needs to be replaced with appropriate coordinates.

The CoDa geometry is described by the ilr transformation representing all pairwise log-ratios; the obtained coordinates form an orthonormal basis in the simplex [1].

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

ilr variables are assigned to an individual compositional part (here to the first one).

Tobit regression is used to estimate censored values:

$$\hat{z}_{i1} = \mathbf{z}_{i,-1}^t \cdot \hat{\beta} + \hat{\sigma} \left[\frac{\phi\left(\frac{\psi_{i1} - \mathbf{z}_{i,-1}^t \cdot \hat{\beta}}{\hat{\sigma}}\right)}{\Phi\left(\frac{\psi_{i1} - \mathbf{z}_{i,-1}^t \cdot \hat{\beta}}{\hat{\sigma}}\right)} \right], \quad (3)$$

where $\hat{\beta}$ are the estimated coefficients, $\hat{\sigma}$ is the estimated standard deviation of the residuals, and ψ_{i1} is the transformed truncation point.

Procedure for imputation of values >UDL

The censored regression algorithm iteratively imputes parts with values above upper detection limit:

1. For imputation in each variable a specific ilr representation is needed.
2. Tobit regression is applied.
3. Values >UDL are replaced by the estimated values using the Tobit regression model.
4. The corresponding inverse ilr transformation is done in order to be in original scale, i.e.

$$x_i = \exp\left(-\sum_{j=1}^{i-1} \frac{1}{\sqrt{(D-j+1)(D-j)}} z_j + \frac{\sqrt{D-i}}{\sqrt{D-i+1}} z_i \right), i = 2, \ldots, D-1. \quad (4)$$

5. Do the same for next variable and recycle the process again.
6. After all parts are imputed, the algorithm starts again until the imputations only change marginally.

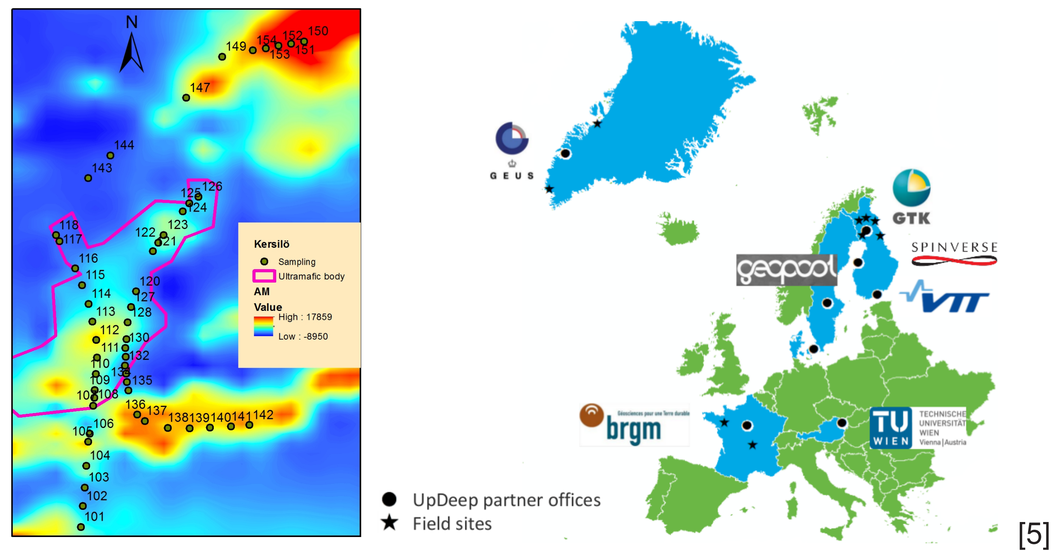
Procedure of simulation study based on real data

Use geochemical data set from Norway with 13 selected variables and 604 observations. Simulation is done using R package `robCompositions` [4]:

1. For each variable in turn, UDL values are generated according to certain quantiles from 0.5 - 0.95.
2. For those quantiles compute imputation for *classical* least-squares and robust regression using an *MM estimator* in order to downweight the influence of large residuals, and for the "naive" approach.
3. Evaluate average effect of all the variables for particular UDLs (quantile). Two measurements of distortion are provided:
   - Relative difference in covariance matrix (RDCM)
   - Compositional error deviation (CED)

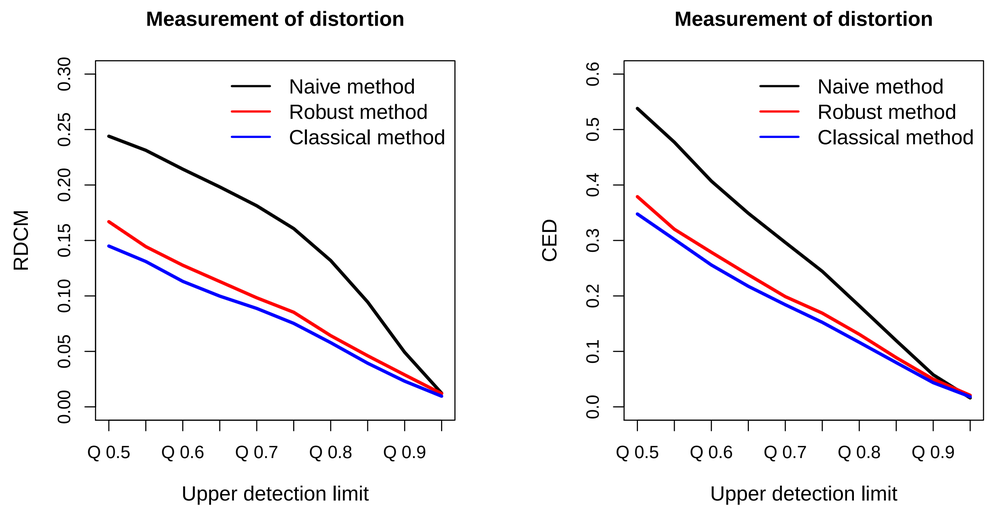
UpDeep project

Upscaling deep buried geochemical exploration techniques into European business [2]. TU Wien is responsible for statistical data analysis.



[5]

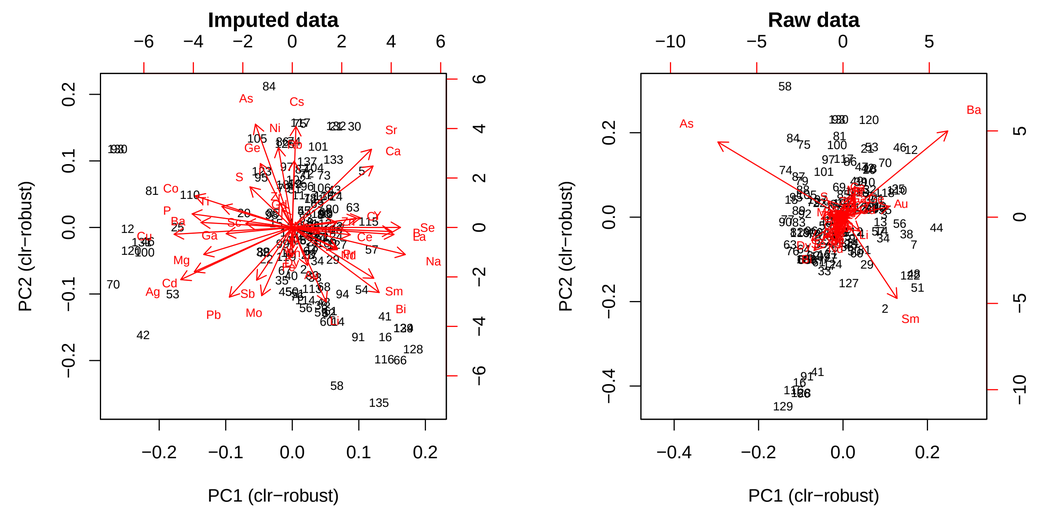
Simulation results



Simulation conclusions:

1. With increasing proportion of values >UDL, the imputation methods based on Tobit regression work better than the naive approach according to both error measurements.
2. Here, the robust method has slightly lower performance than the classical method.

Biplots - real data example



- The imputed data reflect much better the multivariate data structure.

Future work

- Perform simulation studies to compare different imputation techniques in different situations based on various real data sets.
- Implement also partial least squares method.
- Implement function for imputation of values above UDL in an R package.

References

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

[2] KAVA Reference: 16329, UpDeep, Upscaling deep buried geochemical exploration techniques into European business (2017–2020).

[3] J. A. Martin-Fernandez, K. Hron, M. Templ, P. Filzmoser, and J. Palarea-Albaladejo (2012). Model-based replacement of rounded zeros in compositional data: classical and robust approaches. *Computational Statistics & Data Analysis*, **56**(9), 2688–2704.

[4] M. Templ et al. (2018). robCompositions: Robust Estimation for Compositional Data. *R package version 2.0.7.*

[5] UltraLIM-project, Ultra low-impact exploration methods in the subarctic (2013–2015). Funded from Tekes Green Mining Programme.