Replacement of values above an upper detection limit in compositions

Dominika Mikšová
joint work with Peter Filzmoser

TU WIEN

Olomouc, September 6, IAMG conference 2018

This activity has received funding from the European Institute of Innovation and Technology (EIT), a body of the European Union, under the Horizon 2020, the EU Framework Programme for Research and Innovation

Content

Motivation

- Geochemical compositions (concentrations of chemical elements e.g. in plants or soil) are often affected by values **exceeding an upper detection limit** (UDL), besides that also by **rounded zeros** - values below lower detection limit (LDL).
- For Compositional Data (CoDa), only the ratios between the variables (parts) contain the relevant information.
- "Advanced" imputation techniques make use of the multivariate information preserving the data structure: CoDa deals with a specific geometry - **Aitchison geometry**.
- Imputation is necessary for statistical analyses that rely on complete data.
- Simple (naive) approach commonly used in practise: Replacing values above UDL by **1.2 times** the UDL.

CoDa transformation and Tobit regression

- Following the ideas as in the Martin-Fernandez et al. (2012)[1].
- Truncated regression model (τ is truncation point):

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

  where ϕ and Φ are density and distribution function of $N(0,1)$, respectively.
- Initialize values >UDL with naive imputation.
- We use the ilr transformation to deal with compositions:

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

[1] J. A. Martın-Fernández et al. (2012). "Model-based replacement of rounded zeros in compositional data: classical and robust approaches". In: Computational Statistics & Data Analysis 56.9, s. 2688–2704.

- Approach based on **Tobit regression** (used for estimation of censored values):

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

where $\hat{\boldsymbol{\beta}}$ 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 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.
4. The corresponding inverse ilr transformation is done, i.e.

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

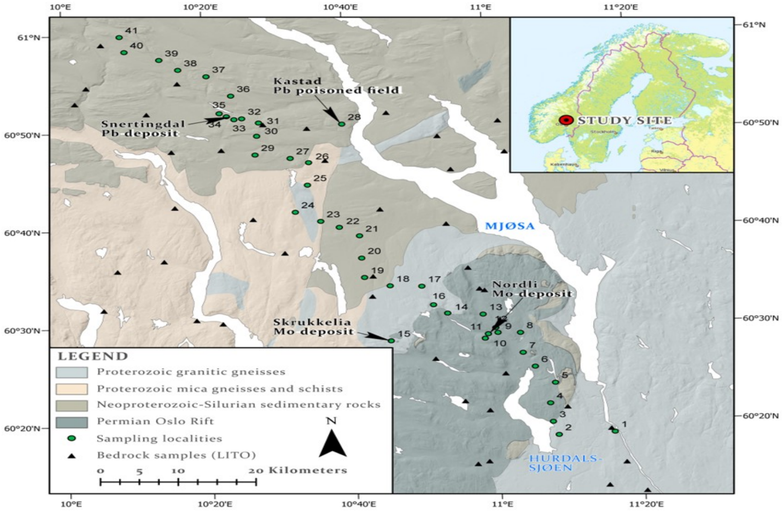
5. Do the same for next variable and recycle the process again.
6. After all parts are imputed, the algorithm starts again until the imputations only change marginally.

Procedure of simulation study based on real data

- We use a geochemical data set from NGU Norway with 30 variables and 604 observations (53 chemical elements analysed).
- The Gjøvik Transect - 100 km long, 41 sample sites, 4 mineralisations crossed[a].
- In total 15 different sample materials (birch, spruce, cowberry, mushroom, O- and C-horizon for soil, etc.) collected at each site.
- For 13 elements analytical quality and the detection limits were sufficient to compare results between all sample media.

[a] Clemens Reimann et al. (2018). "The response of 12 different plant materials and one mushroom to Mo and Pb mineralization along a 100-km transect in southern central Norway". In: Geochemistry: Explor., Envir., Anal. DOI: 10.1144/geochem2017-089.

Procedure of simulation study based on real data



Obrázek: Bedrock geological map showing the Gjøvik transect sample sites[a]

Simulation is done using R package `robCompositions`:

1st scenario:

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 and robust regression (downweight outliers), and for the "naive" approach.
3. Evaluate average effect of all the variables for particular UDLs (quantile) - two measurements of distortion are used:

Simulation based on real data

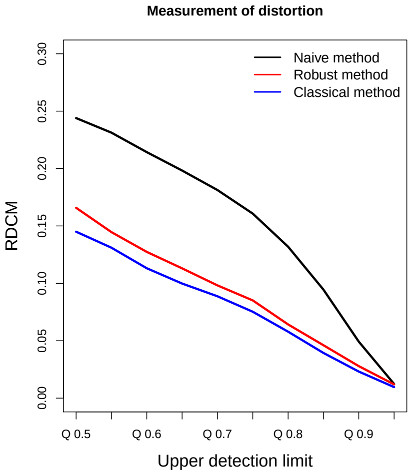
- Relative difference in covariance matrix (RDCM). The sample covariance matrices are computed with the same ilr transformed observations.

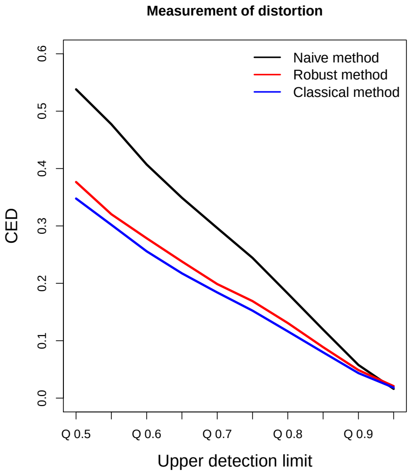
$$\frac{\|\mathbf{S}-\mathbf{S}^*\|_F}{\|\mathbf{S}\|_F} = \frac{\sqrt{\sum_{i,j=1}^{D-1}(s_{ij}-s_{ij}^*)^2}}{\sqrt{\sum_{i,j=1}^{D-1}s_{ij}^2}} \tag{5}$$

- Compositional error deviation (CED). Normalized Aitchison distance between two data sets. M is index for samples containing at least one value >UDL.

$$\frac{\frac{1}{n_M}\sum_{k\in M} d_a(\mathbf{x}_k,\mathbf{x}_k^*)}{\max_{\{\mathbf{x}_i,\mathbf{x}_j\in\mathbf{X}\}}\{d_a(\mathbf{x}_i,\mathbf{x}_j)\}} \tag{6}$$

Simulation based on real data





Original data - Artificial UDL = 2309 (quantile 85%)

	Al	Ba	Cd	Ce	Co	Cs	Cu	Fe	Mn	Mo	Na	Ni	Zn
15	23.7	137.48	0.861	0.048	0.589	0.0130	4.42	88	3942	0.0618	10.8	1.167	254.3
16	26.8	72.62	0.263	0.055	0.802	0.0139	3.61	106	3700	0.1296	12.6	0.942	195.6
17	25.6	94.50	0.246	0.037	0.646	0.0117	3.72	101	1855	0.0516	13.1	4.433	167.2
18	31.6	151.95	0.556	0.090	0.673	0.0263	4.33	122	3168	0.0927	21.6	1.069	273.4
19	32.2	107.70	0.459	0.116	0.120	0.0608	3.71	84	4011	0.1310	49.1	1.260	262.9
20	37.1	106.42	0.523	0.091	0.145	0.1870	3.54	89	3281	0.1700	24.3	1.096	311.8
21	43.8	135.49	0.333	0.113	0.326	0.0269	4.67	124	1616	0.0457	32.4	4.827	294.7
22	42.4	140.18	0.634	0.090	0.096	0.2613	4.99	110	5392	0.3471	107.7	1.581	310.4
23	46.7	186.99	0.970	0.150	0.087	0.0285	3.55	124	3543	0.1669	26.4	0.515	653.3
24	21.1	153.52	0.274	0.060	0.268	0.0855	3.59	121	1689	0.1233	16.5	0.459	203.2
25	38.7	120.68	0.504	0.079	1.190	0.0763	4.30	106	2171	0.0925	23.7	1.758	313.8
26	42.1	145.07	0.440	0.068	1.246	0.1553	4.52	104	3584	0.0880	18.9	2.166	314.4
27	69.0	95.48	0.528	0.179	0.783	0.1957	5.22	176	3405	0.2267	45.6	2.219	184.1
28	54.6	101.54	0.500	0.199	0.342	0.1471	5.50	111	3151	0.1259	30.1	3.095	301.9
29	41.9	107.19	0.479	0.116	1.225	0.0401	4.26	118	2720	0.1095	28.7	1.471	257.9
30	32.2	37.67	0.502	0.057	0.330	0.0795	4.00	96	2045	0.0759	36.2	1.222	246.3
31	40.1	61.76	0.590	0.186	0.319	0.0266	4.83	101	4737	0.1348	40.2	1.032	323.5
32	21.4	49.66	0.501	0.053	0.575	0.0429	3.06	83	3097	0.0761	12.9	0.792	242.7
33	19.0	49.25	0.345	0.040	0.575	0.0328	3.24	117	1537	0.1573	13.1	1.525	144.4

Replacement by Inf

	Al	Ba	Cd	Ce	Co	Cs	Cu	Fe	Mn	Mo	Na	Ni	Zn
15	23.7	137.48	0.861	0.048	0.589	0.0130	4.42	88	Inf	0.0618	10.8	1.167	254.3
16	26.8	72.62	0.263	0.055	0.802	0.0139	3.61	106	Inf	0.1296	12.6	0.942	195.6
17	25.6	94.50	0.246	0.037	0.646	0.0117	3.72	101	1855	0.0516	13.1	4.433	167.2
18	31.6	151.95	0.556	0.090	0.673	0.0263	4.33	122	Inf	0.0927	21.6	1.069	273.4
19	32.2	107.70	0.459	0.116	0.120	0.0608	3.71	84	Inf	0.1310	49.1	1.260	262.9
20	37.1	106.42	0.523	0.091	0.145	0.1870	3.54	89	Inf	0.1700	24.3	1.096	311.8
21	43.8	135.49	0.333	0.113	0.326	0.0269	4.67	124	1616	0.0457	32.4	4.827	294.7
22	42.4	140.18	0.634	0.090	0.096	0.2613	4.99	110	Inf	0.3471	107.7	1.581	310.4
23	46.7	186.99	0.970	0.150	0.087	0.0285	3.55	124	Inf	0.1669	26.4	0.515	653.3
24	21.1	153.52	0.274	0.060	0.268	0.0855	3.59	121	1689	0.1233	16.5	0.459	203.2
25	38.7	120.68	0.504	0.079	1.190	0.0763	4.30	106	2171	0.0925	23.7	1.758	313.8
26	42.1	145.07	0.440	0.068	1.246	0.1553	4.52	104	Inf	0.0880	18.9	2.166	314.4
27	69.0	95.48	0.528	0.179	0.783	0.1957	5.22	176	Inf	0.2267	45.6	2.219	184.1
28	54.6	101.54	0.500	0.199	0.342	0.1471	5.50	111	Inf	0.1259	30.1	3.095	301.9
29	41.9	107.19	0.479	0.116	1.225	0.0401	4.26	118	Inf	0.1095	28.7	1.471	257.9
30	32.2	37.67	0.502	0.057	0.330	0.0795	4.00	96	2045	0.0759	36.2	1.222	246.3
31	40.1	61.76	0.590	0.186	0.319	0.0266	4.83	101	Inf	0.1348	40.2	1.032	323.5
32	21.4	49.66	0.501	0.053	0.575	0.0429	3.06	83	Inf	0.0761	12.9	0.792	242.7
33	19.0	49.25	0.345	0.040	0.575	0.0328	3.24	117	1537	0.1573	13.1	1.525	144.4

Imputed data based on Tobit regression

	Al	Ba	Cd	Ce	Co	Cs	Cu	Fe	Mn	Mo	Na	Ni	Zn
15	23.7	137.48	0.861	0.048	0.589	0.0130	4.42	88	3392.272	0.0618	10.8	1.167	254.3
16	26.8	72.62	0.263	0.055	0.802	0.0139	3.61	106	3299.623	0.1296	12.6	0.942	195.6
17	25.6	94.50	0.246	0.037	0.646	0.0117	3.72	101	1855.000	0.0516	13.1	4.433	167.2
18	31.6	151.95	0.556	0.090	0.673	0.0263	4.33	122	3540.451	0.0927	21.6	1.069	273.4
19	32.2	107.70	0.459	0.116	0.120	0.0608	3.71	84	4016.379	0.1310	49.1	1.260	262.9
20	37.1	106.42	0.523	0.091	0.145	0.1870	3.54	89	4395.477	0.1700	24.3	1.096	311.8
21	43.8	135.49	0.333	0.113	0.326	0.0269	4.67	124	1616.000	0.0457	32.4	4.827	294.7
22	42.4	140.18	0.634	0.090	0.096	0.2613	4.99	110	5188.908	0.3471	107.7	1.581	310.4
23	46.7	186.99	0.970	0.150	0.087	0.0285	3.55	124	4517.841	0.1669	26.4	0.515	653.3
24	21.1	153.52	0.274	0.060	0.268	0.0855	3.59	121	1689.000	0.1233	16.5	0.459	203.2
25	38.7	120.68	0.504	0.079	1.190	0.0763	4.30	106	2171.000	0.0925	23.7	1.758	313.8
26	42.1	145.07	0.440	0.068	1.246	0.1553	4.52	104	4387.731	0.0880	18.9	2.166	314.4
27	69.0	95.48	0.528	0.179	0.783	0.1957	5.22	176	3600.537	0.2267	45.6	2.219	184.1
28	54.6	101.54	0.500	0.199	0.342	0.1471	5.50	111	4182.040	0.1259	30.1	3.095	301.9
29	41.9	107.19	0.479	0.116	1.225	0.0401	4.26	118	3536.009	0.1095	28.7	1.471	257.9
30	32.2	37.67	0.502	0.057	0.330	0.0795	4.00	96	2045.000	0.0759	36.2	1.222	246.3
31	40.1	61.76	0.590	0.186	0.319	0.0266	4.83	101	3371.816	0.1348	40.2	1.032	323.5
32	21.4	49.66	0.501	0.053	0.575	0.0429	3.06	83	3238.602	0.0761	12.9	0.792	242.7
33	19.0	49.25	0.345	0.040	0.575	0.0328	3.24	117	1537.000	0.1573	13.1	1.525	144.4

Imputed data by naive approach

	Al	Ba	Cd	Ce	Co	Cs	Cu	Fe	Mn	Mo	Na	Ni	Zn
15	23.7	137.48	0.861	0.048	0.589	0.0130	4.42	88	2770.8	0.0618	10.8	1.167	254.3
16	26.8	72.62	0.263	0.055	0.802	0.0139	3.61	106	2770.8	0.1296	12.6	0.942	195.6
17	25.6	94.50	0.246	0.037	0.646	0.0117	3.72	101	1855.0	0.0516	13.1	4.433	167.2
18	31.6	151.95	0.556	0.090	0.673	0.0263	4.33	122	2770.8	0.0927	21.6	1.069	273.4
19	32.2	107.70	0.459	0.116	0.120	0.0608	3.71	84	2770.8	0.1310	49.1	1.260	262.9
20	37.1	106.42	0.523	0.091	0.145	0.1870	3.54	89	2770.8	0.1700	24.3	1.096	311.8
21	43.8	135.49	0.333	0.113	0.326	0.0269	4.67	124	1616.0	0.0457	32.4	4.827	294.7
22	42.4	140.18	0.634	0.090	0.096	0.2613	4.99	110	2770.8	0.3471	107.7	1.581	310.4
23	46.7	186.99	0.970	0.150	0.087	0.0285	3.55	124	2770.8	0.1669	26.4	0.515	653.3
24	21.1	153.52	0.274	0.060	0.268	0.0855	3.59	121	1689.0	0.1233	16.5	0.459	203.2
25	38.7	120.68	0.504	0.079	1.190	0.0763	4.30	106	2171.0	0.0925	23.7	1.758	313.8
26	42.1	145.07	0.440	0.068	1.246	0.1553	4.52	104	2770.8	0.0880	18.9	2.166	314.4
27	69.0	95.48	0.528	0.179	0.783	0.1957	5.22	176	2770.8	0.2267	45.6	2.219	184.1
28	54.6	101.54	0.500	0.199	0.342	0.1471	5.50	111	2770.8	0.1259	30.1	3.095	301.9
29	41.9	107.19	0.479	0.116	1.225	0.0401	4.26	118	2770.8	0.1095	28.7	1.471	257.9
30	32.2	37.67	0.502	0.057	0.330	0.0795	4.00	96	2045.0	0.0759	36.2	1.222	246.3
31	40.1	61.76	0.590	0.186	0.319	0.0266	4.83	101	2770.8	0.1348	40.2	1.032	323.5
32	21.4	49.66	0.501	0.053	0.575	0.0429	3.06	83	2770.8	0.0761	12.9	0.792	242.7
33	19.0	49.25	0.345	0.040	0.575	0.0328	3.24	117	1537.0	0.1573	13.1	1.525	144.4

Comparison

Tobit reg. → ← "naive"

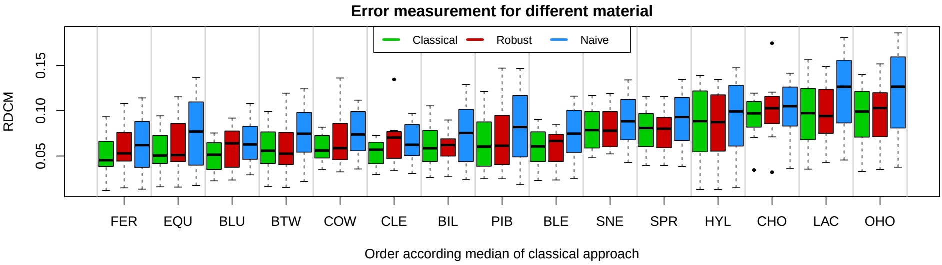
	Mn	Mn (Tobit reg.)	Mn ("naive")	Mn	
15	3942	3392.272	2770.8	3942	15
16	3700	3299.623	2770.8	3700	16
17	1855	1855.000	1855.0	1855	17
18	3168	3540.451	2770.8	3168	18
19	4011	4016.379	2770.8	4011	19
20	3281	4395.477	2770.8	3281	20
21	1616	1616.000	1616.0	1616	21
22	5392	5188.908	2770.8	5392	22
23	3543	4517.841	2770.8	3543	23
24	1689	1689.000	1689.0	1689	24
25	2171	2171.000	2171.0	2171	25
26	3584	4387.731	2770.8	3584	26
27	3405	3600.537	2770.8	3405	27
28	3151	4182.040	2770.8	3151	28
29	2720	3536.009	2770.8	2720	29
30	2045	2045.000	2045.0	2045	30
31	4737	3371.816	2770.8	4737	31
32	3097	3238.602	2770.8	3097	32
33	1537	1537.000	1537.0	1537	33

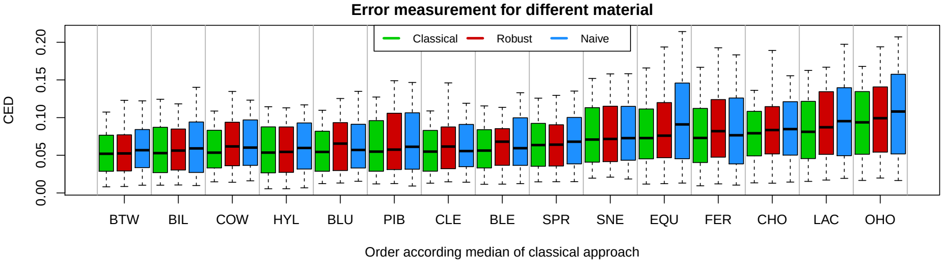
Simulation based on real data

2nd scenario:

- Boxplots for different material separately $\rightarrow$ plants, soil (CHO; OHO) and fungi (LAC).
- 13 variables selected.
- Two measurements of distortion are provided.
- Sorted according median of classical approach.

Simulation based on real data



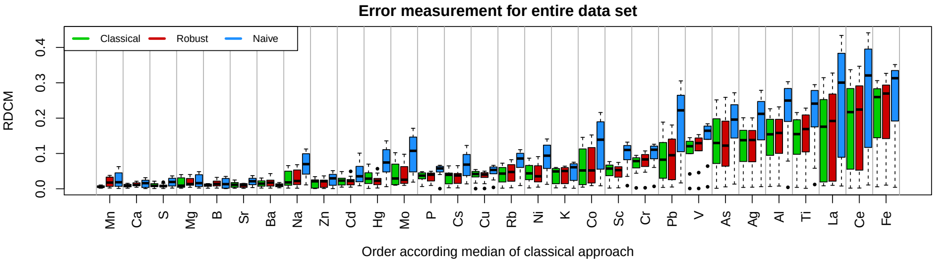


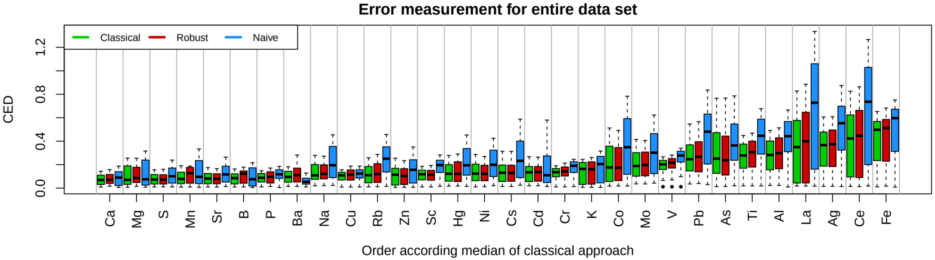
Simulation based on real data

3rd scenario:

- Boxplots for 30 different elements.
- Using entire data set.
- Sorted according median of classical approach.
- Results for RDCM: Naive approach is better than classical method just for element **Ba** and better than robust method for **B, Ba, Mn**, otherwise it gives clear advantage of imputation based on Tobit regression.

Simulation based on real data



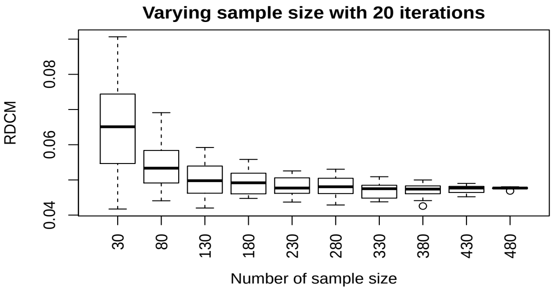


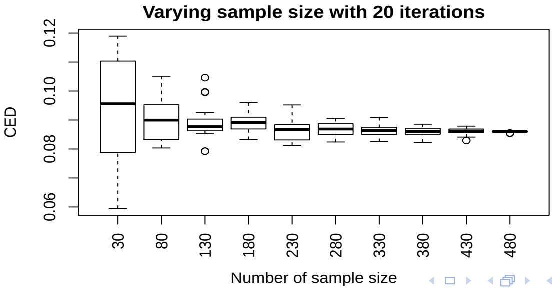
Simulation based on real data

4th scenario:

- Varying sample size with 20 iterations.
- Using just plant material (birch, spruce, cowberry, etc.)
- Computed imputation method for classical approach.
- Significant difference of error measurements for small data and the big one.

Simulation based on real data





Key points of the presentation:

- When applying multivariate statistical methods, it is necessary to have a complete data set available.
- Classical and robust methods are applied after ilr transformation.
- With increasing proportion of values >UDL, the imputation method based on Tobit regression performed better than the naive aproach according to both error measurements.
- Imputation produces minor distortion in the covariance structure of tha data.
- Next step: combined strategy for replacement of both UDL and LDL values.

Thank you for your attention!

References:

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.

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.

C. Reimann, P. Englmaier, B. Flem, O. A. Eggen, T. E. Finne, M. Andersson P. Filzmoser (2018). The response of 12 different plant materials and one mushroom to Mo and Pb mineralization along a 100-km transect in southern central Norway. Geochemistry: Exploration, Environment, Analysis, **18**(3), 204–215.

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