The statistical analysis of compositional data: Exploratory analysis

Prof. Dr. Vera Pawlowsky-Glahn
Prof. Dr. Juan José Egozcue
Ass. Prof. Dr. René Meziat

Instituto Colombiano del Petróleo
Piedecuesta, Santander, Colombia
March 20–23, 2007

the starting point

It is sometimes considered a paradox that the answer depends not only on the observations but on the question: it should be a platitude.

Harold Jeffreys

- the starting point is (or should be) always a question
- the second step is proper sampling to answer the question
- the next step is checking for zeros in the data
- then one can proceed with exploratory analysis

the starting point

It is sometimes considered a paradox that the answer depends not only on the observations but on the question: it should be a platitude.

Harold Jeffreys

- the starting point is (or should be) always a question
- the second step is proper sampling to answer the question
- the next step is checking for zeros in the data
- then one can proceed with exploratory analysis

the treatment of zeros

case 1: the part with zeros is not important for the study
$\Rightarrow$ the part should be omitted

case 2: the part is important, the zeros are essential
$\Rightarrow$ divide the sample into two or more populations, according to the presence/absence of zeros

case 3: the part is important, the zeros are rounded zeros
$\Rightarrow$ use imputation techniques

centre and total variance

centre of a compositional dataset of size n
= **closed geometric mean**

$$\mathbf{g} = \mathcal{C}\left[g_1, g_2, \ldots, g_D\right], \text{ with } g_i = \left(\prod_{j=1}^{n} x_{ji}\right)^{1/n}$$

total variance = measure of total dispersion

$$\mathrm{totvar}[\mathbf{X}] = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\mathrm{var}\left[\ln\frac{x_i}{x_j}\right]$$

other summary statistics: **min, max, Q1, median, Q3**

example: Kilauea Iki data

Compositional Descriptive Statistics

	SiO2	TiO2	Al2O3	Fe2O3	FeO	MnO	MgO	CaO	Na2O	K2O	P2O5
Center	**48.57**	**2.35**	**11.23**	**1.84**	**9.91**	**0.18**	**13.74**	**9.65**	**1.82**	**0.48**	**0.22**
Min	45.58	1.54	8.18	1.04	8.92	0.17	8.85	6.80	1.28	0.31	0.15
Max	49.68	3.31	12.93	2.81	10.46	0.18	23.91	11.06	2.25	0.80	0.30
Q25	47.47	2.13	10.54	1.49	9.45	0.17	10.81	9.09	1.63	0.40	0.20
Median	48.45	2.35	11.60	1.83	9.91	0.18	13.24	9.88	1.89	0.46	0.23
Q75	49.07	2.49	12.23	2.21	10.15	0.18	16.49	10.60	2.02	0.56	0.24

total variance: 0.3163

check for unreasonable values, due to errors or outliers

variation matrix

$$\mathbf{T} = \begin{pmatrix} \mathrm{var}\left[\ln \frac{x_1}{x_1}\right] & \mathrm{var}\left[\ln \frac{x_1}{x_2}\right] & \cdots & \mathrm{var}\left[\ln \frac{x_1}{x_D}\right] \\ \mathrm{var}\left[\ln \frac{x_2}{x_1}\right] & \mathrm{var}\left[\ln \frac{x_2}{x_2}\right] & \cdots & \mathrm{var}\left[\ln \frac{x_2}{x_D}\right] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{var}\left[\ln \frac{x_D}{x_1}\right] & \mathrm{var}\left[\ln \frac{x_D}{x_2}\right] & \cdots & \mathrm{var}\left[\ln \frac{x_D}{x_D}\right] \end{pmatrix}$$

- **T** is symmetric
- **T** has zeros in its diagonal
- neither the total variance nor any single entry in **T** depend on the closure constant $\kappa \Rightarrow$ rescaling has no effect

variation array

$$\begin{pmatrix} - & \mathrm{var}\left[\ln \frac{x_1}{x_2}\right] & \cdots & \mathrm{var}\left[\ln \frac{x_1}{x_D}\right] \\ \mathrm{E}\left[\ln \frac{x_1}{x_2}\right] & - & \ddots & \vdots \\ \vdots & \ddots & - & \mathrm{var}\left[\ln \frac{x_{D-1}}{x_D}\right] \\ \mathrm{E}\left[\ln \frac{x_1}{x_D}\right] & \cdots & \mathrm{E}\left[\ln \frac{x_{D-1}}{x_D}\right] & - \end{pmatrix}$$

- upper triangle: variances of simple logratios
- lower triangle: means of simple logratios

example: Kilauea Iki data

Compositional Descriptive Statistics

	SiO2	TiO2	Al2O3	Fe2O3	FeO	MnO	MgO	CaO	Na2O	K2O	P2O5
Center	**48.57**	**2.35**	**11.23**	**1.84**	**9.91**	**0.18**	**13.74**	**9.65**	**1.82**	**0.48**	**0.22**
Min	45.58	1.54	8.18	1.04	8.92	0.17	8.85	6.80	1.28	0.31	0.15
Max	49.68	3.31	12.93	2.81	10.46	0.18	23.91	11.06	2.25	0.80	0.30
Q25	47.47	2.13	10.54	1.49	9.45	0.17	10.81	9.09	1.63	0.40	0.20
Median	48.45	2.35	11.60	1.83	9.91	0.18	13.24	9.88	1.89	0.46	0.23
Q75	49.07	2.49	12.23	2.21	10.15	0.18	16.49	10.60	2.02	0.56	0.24

Variation Array

	SiO2	TiO2	Al2O3	Fe2O3	FeO	MnO	MgO	CaO	Na2O	K2O	P2O5
SiO2		0.026	0.012	0.077	0.003	0.001	0.098	0.015	0.019	0.062	0.023
TiO2	3.027		0.007	0.123	0.040	0.035	0.218	0.010	0.004	0.031	0.001
Al2O3	1.465	-1.562		0.106	0.023	0.018	0.178	0.001	0.004	0.037	0.005
Fe2O3	3.274	0.247	1.810		0.094	0.073	0.114	0.114	0.105	0.197	0.125
FeO	1.589	-1.438	0.125	-1.685		0.003	0.081	0.026	0.032	0.073	0.036
MnO	5.604	2.577	4.140	2.330	4.015		0.085	0.019	0.026	0.071	0.032
MgO	1.262	-1.764	-0.202	-2.012	-0.327	-4.342		0.184	0.194	0.275	0.212
CaO	1.616	-1.411	0.152	-1.658	0.027	-3.988	0.354		0.007	0.035	0.008
Na2O	3.282	0.256	1.818	0.008	1.693	-2.322	2.020	1.666		0.050	0.004
K2O	4.617	1.590	3.152	1.342	3.028	-0.988	3.354	3.001	1.334		0.030
P2O5	5.391	2.364	3.926	2.116	3.801	-0.214	4.128	3.774	2.108	0.774	

Variances

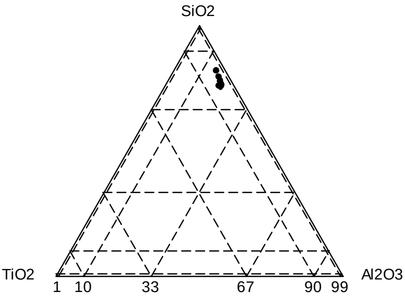
Means

centring a dataset

consider a sample $\mathbf{X}$ with centre $\mathbf{g}$,
compute $\mathbf{X} \ominus \mathbf{g} = \hat{\mathbf{X}}$, plot $\hat{\mathbf{X}}$
notes:

- the centred sample $\hat{\mathbf{X}}$ will gravitate around the barycentre
- helpful to visualise data in a ternary diagram
- variation array and total variance do not change
- perturbation transforms straight lines into straight lines $\Rightarrow$ gridlines and compositional fields can be included in the graphical representation without the risk of a nonlinear distortion

example: Kilauea Iki data

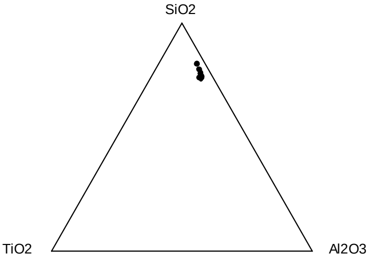


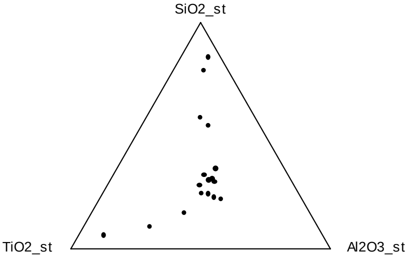
standardisation

consider a sample $\mathbf{X}$ with centre $\mathbf{g}$ and total variance s_T^2,
compute $(1/s_T) \odot \mathbf{X} \ominus \mathbf{g} = \mathbf{X}_s$, plot $\mathbf{X}_s$
notes:

- the standardised sample $\mathbf{X}_s$ will gravitate around the barycentre and will have unit total variance
- helpful to visualise data in a ternary diagram
- variation array and total variance change (lower triangle is identically zero; logratio variances are different)

example: Kilauea Iki data





the compositional biplot: a graphical display

a biplot represents simultaneously the rows and columns of any matrix by means of a rank-2 approximation

construction:

- consider the centred matrix $\mathbf{X} \ominus \mathbf{g} = \hat{\mathbf{X}}$
- compute $\mathbf{Z} = \mathrm{clr}(\hat{\mathbf{X}})$ and its singular value decomposition

$$\mathbf{Z} = \mathbf{L}\mathrm{diag}(k_1, k_2, \ldots, k_s)\mathbf{M}', \quad k_i = \sqrt{\lambda_i}$$

with $\lambda_1 > \cdots > \lambda_s$ the s positive eigenvalues of $\mathbf{ZZ}'$ or $\mathbf{Z}'\mathbf{Z}$, $\mathbf{L}$ the eigenvectors of $\mathbf{ZZ}'$, $\mathbf{M}$ the eigenvectors of $\mathbf{Z}'\mathbf{Z}$

the compositional biplot: a graphical display

the best rank-2 approximation in the least squares sense is

$$\mathbf{Y} = \begin{pmatrix} \ell_{11} & \ell_{21} \\ \ell_{12} & \ell_{22} \\ \vdots & \vdots \\ \ell_{1n} & \ell_{2n} \end{pmatrix} \begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix} \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1D} \\ m_{21} & m_{22} & \cdots & m_{2D} \end{pmatrix}$$

with a proportion of variability retained $= \dfrac{\lambda_1 + \lambda_2}{\sum_{i=1}^{s} \lambda_i}$

the compositional biplot: a graphical display

representation: write $\mathbf{Y} = \mathbf{G}\mathbf{H}'$ with

$$\mathbf{G} = \begin{pmatrix} \sqrt{n-1}\ell_{11} & \sqrt{n-1}\ell_{21} \\ \sqrt{n-1}\ell_{12} & \sqrt{n-1}\ell_{22} \\ \vdots & \vdots \\ \sqrt{n-1}\ell_{1n} & \sqrt{n-1}\ell_{2n} \end{pmatrix} = \begin{pmatrix} \mathbf{g}_1 \\ \mathbf{g}_2 \\ \vdots \\ \mathbf{g}_n \end{pmatrix}$$

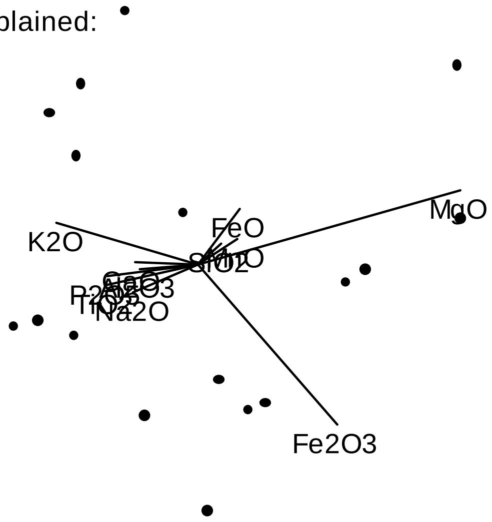
$$\mathbf{H}' = \begin{pmatrix} \frac{k_1 m_{11}}{\sqrt{n-1}} & \frac{k_1 m_{12}}{\sqrt{n-1}} & \cdots & \frac{k_1 m_{1D}}{\sqrt{n-1}} \\ \frac{k_2 m_{21}}{\sqrt{n-1}} & \frac{k_2 m_{22}}{\sqrt{n-1}} & \cdots & \frac{k_2 m_{2D}}{\sqrt{n-1}} \end{pmatrix} = \begin{pmatrix} \mathbf{h}'_1 & \mathbf{h}'_2 & \cdots & \mathbf{h}'_D \end{pmatrix}$$

plot $\mathbf{g}_1$, $\mathbf{g}_2$, ..., $\mathbf{g}_n$ = row markers = projections of samples
plot $\mathbf{h}_1$, $\mathbf{h}_2$, ..., $\mathbf{h}_D$ = column markers = projections of clr-parts

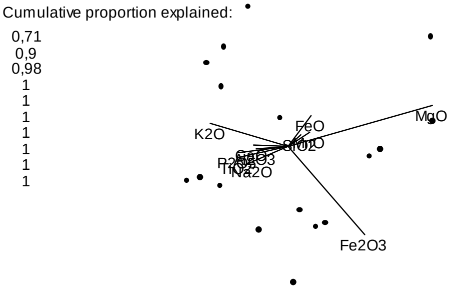
example: Kilauea Iki data

Cumulative proportion explained:

0,71
0,9
0,98
1
1
1
1
1
1
1

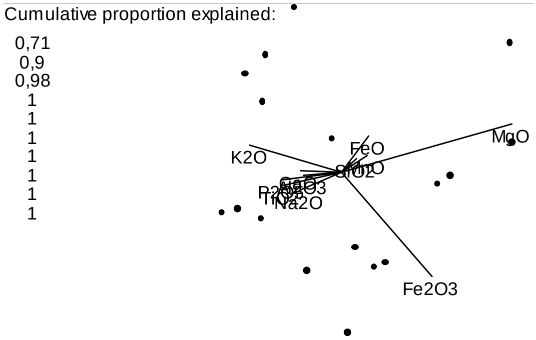


interpretation of a compositional biplot (1)



origin O:	center of the dataset
vertices $\mathbf{h}_j$:	one for each part
case markers $\mathbf{g}_i$:	one for each sample
rays $\overline{Oj}$:	join of O to a vertex j;
links $\overline{jk}$:	join of two vertices j and k

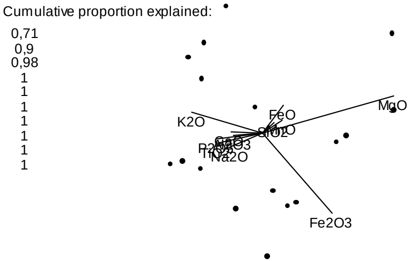
interpretation of a compositional biplot (2)



$$|\overline{jk}|^2 \approx \mathrm{var}\left[\ln \frac{x_j}{x_k}\right] ; \qquad |\overline{Oj}|^2 \approx \mathrm{var}\left[\ln \frac{x_j}{g(x)}\right]$$

subcompositional analysis: select vertices

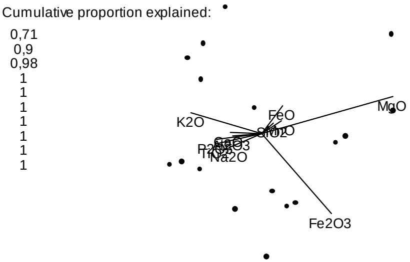
interpretation of a compositional biplot (3)

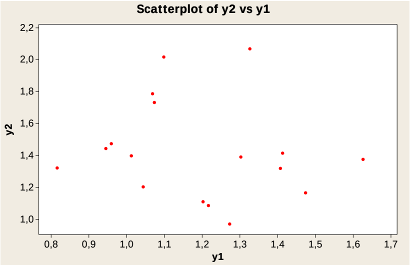


- $\overline{jk}$ and $\overline{i\ell}$ intersect in $M \Rightarrow \cos(jMi) \approx \mathrm{corr}\left[\ln \frac{x_j}{x_k}, \ln \frac{x_i}{x_\ell}\right]$
- $\overline{jk}$ and $\overline{i\ell}$ at a **right angle** $\Rightarrow$ **possible zero correlation**

$$\cos(jMi) \approx 0 \Rightarrow \mathrm{corr}\left[\ln \frac{x_j}{x_k}, \ln \frac{x_i}{x_\ell}\right] \approx 0$$

example: Kilauea Iki data



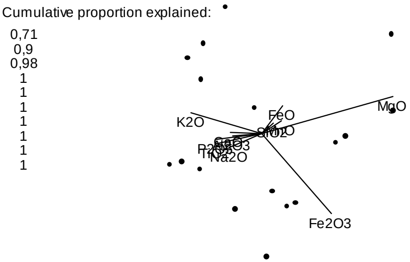


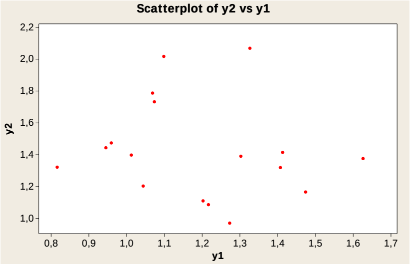
$$\overline{\mathrm{FeO}\,\mathrm{Fe_2O_3}} \perp \overline{\mathrm{MgO}\,\mathrm{Na_2O}} \Rightarrow \text{possibly}$$

$$\mathrm{corr}\left[\frac{1}{\sqrt{2}}\ln\frac{\mathrm{FeO}}{\mathrm{Fe_2O_3}}, \frac{1}{\sqrt{2}}\ln\frac{\mathrm{MgO}}{\mathrm{Na_2O}}\right] \approx 0$$

in fact: Pearson correlation = -0,149; P-Value = 0,567

example: Kilauea Iki data





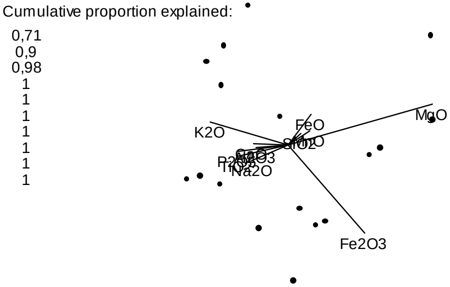
$\overline{\mathrm{FeO}\,\mathrm{Fe_2O_3}} \perp \overline{\mathrm{MgO}\,\mathrm{Na_2O}} \Rightarrow$ possibly

$$\mathrm{corr}\left[\frac{1}{\sqrt{2}}\ln\frac{\mathrm{FeO}}{\mathrm{Fe_2O_3}}, \frac{1}{\sqrt{2}}\ln\frac{\mathrm{MgO}}{\mathrm{Na_2O}}\right] \approx 0$$

in fact: Pearson correlation = -0,149; P-Value = 0,567

V. Pawlowsky-Glahn and J. J. Egozcue

interpretation of a compositional biplot (4)

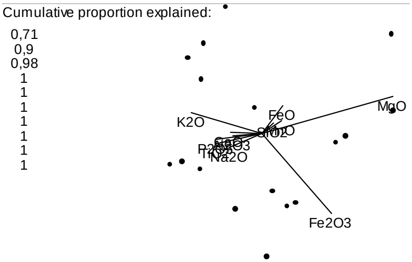


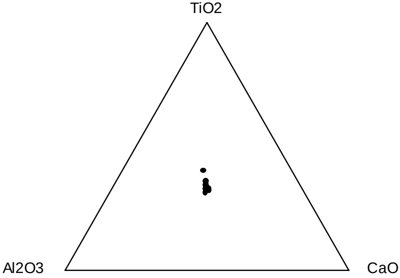
- **coincident vertices** $\Rightarrow x_j, x_k \approx$ **redundant**:

$$\overline{jk} \approx 0 \Rightarrow \mathrm{var}\left[\ln \frac{x_j}{x_k}\right] \approx 0 \Rightarrow \frac{x_j}{x_k} \approx \textbf{constant}$$

- **collinear vertices** $\Rightarrow$ **one-dimensional variability**
  $\Rightarrow$ compositions plot along a compositional line

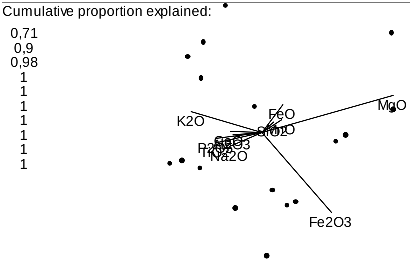
example: Kilauea Iki data

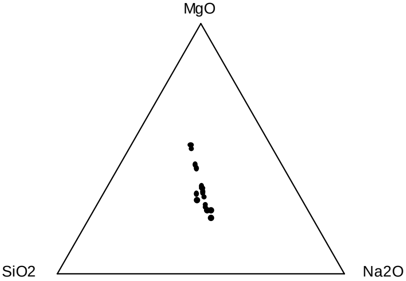




coincident vertices: TiO_2, Al_2O_3, CaO, Na_2O, P_2O_5, CO_2 $\Rightarrow$ redundant information

example: Kilauea Iki data





collinear vertices: MgO, SiO_2, Na_2O $\Rightarrow$ **compositional line**

building an orthonormal basis

- **strategy:** define a sequential binary partition (SBP)
- **criteria:**
  - expert knowledge
  - new theory related to the question
  - previous steps (variation matrix, biplot)

SBP for Kilauea Iki data

order	FeO	Fe_2O_3	MgO	SiO_2	TiO_2	Al_2O_3	CaO	Na_2O	K_2O	P_2O_5	MnO
1	+1	+1	-1	-1	-1	-1	-1	-1	-1	-1	-1
2	+1	-1	0	0	0	0	0	0	0	0	0
3	0	0	+1	-1	-1	-1	-1	-1	-1	-1	-1
4	0	0	0	+1	-1	-1	-1	-1	-1	-1	-1
5	0	0	0	0	+1	-1	-1	-1	-1	-1	-1
6	0	0	0	0	0	+1	-1	-1	-1	-1	-1
7	0	0	0	0	0	0	+1	-1	-1	-1	-1
8	0	0	0	0	0	0	0	+1	-1	-1	-1
9	0	0	0	0	0	0	0	0	+1	-1	-1
10	0	0	0	0	0	0	0	0	0	+1	-1

- balance 1: separates $\{FeO, Fe_2O_3\}$ from the other parts
- balance 2: separates FeO from Fe_2O_3
- balance 3: separates MgO from the remaining parts
- balance 4: separates SiO_2 from the remaining parts
- etc.

summary statistics for balances of Kilauea Iki data

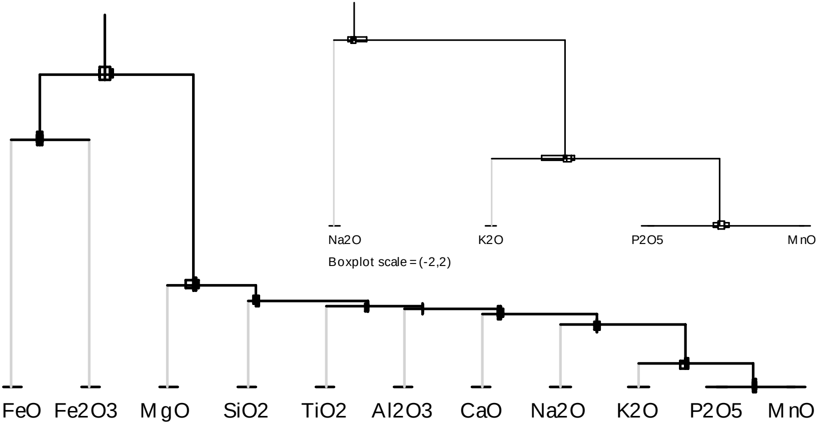
Data in the sample 17
Total Variance 0.298

Number of parts 11
Computed balances 10

coord.num.	x1*	x2*	x3*	x4*	x5*	x6*	x7*	x8*	x9*	x10*
mean coord	0.622	1.192	1.756	3.341	0.588	2.408	2.779	1.664	0.719	0.151
S^2 centre	0.707	0.844	0.923	0.991	0.697	0.968	0.981	0.913	0.734	0.553
std. coord	0.202	0.210	0.379	0.107	0.063	0.034	0.055	0.092	0.163	0.122
var. coord	0.041	0.044	0.143	0.011	0.004	0.001	0.003	0.008	0.026	0.015
min. coord	0.279	0.816	1.219	3.197	0.489	2.336	2.608	1.431	0.541	-0.089
max. coord	0.946	1.626	2.558	3.591	0.771	2.459	2.868	1.761	1.108	0.402

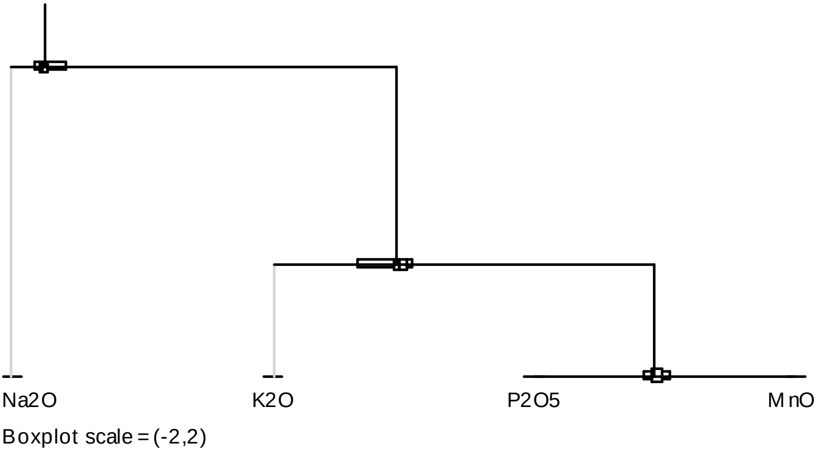
x1*	0.2787	0.2787	0.3239	0.3850	0.6117	0.7967	0.8280	0.8326	0.8895
x2*	0.8160	0.8160	0.8804	1.0124	1.1503	1.3143	1.4130	1.4739	1.5500
x3*	1.2188	1.2188	1.2885	1.4505	1.6740	1.8297	2.2106	2.4925	2.5255
x4*	3.1968	3.1968	3.2114	3.2447	3.3148	3.3536	3.4485	3.5554	3.5730
x5*	0.4889	0.4889	0.5064	0.5524	0.5768	0.5881	0.6242	0.7054	0.7384
x6*	2.3360	2.3360	2.3379	2.3944	2.4176	2.4246	2.4408	2.4419	2.4503
x7*	2.6081	2.6081	2.6715	2.7490	2.7906	2.8029	2.8193	2.8419	2.8548
x8*	1.4314	1.4314	1.4388	1.6164	1.6951	1.7128	1.7401	1.7503	1.7554
x9*	0.5412	0.5412	0.5414	0.6092	0.6847	0.7353	0.8632	1.1005	1.1040
x10*	-0.0885	-0.0885	-0.0657	0.0382	0.1733	0.1884	0.2600	0.3271	0.3644
Prob.	0.0100	0.0500	0.1000	0.2500	0.5000	0.7500	0.9000	0.9500	0.9900

balance-dendrogram for Kilauea Iki data

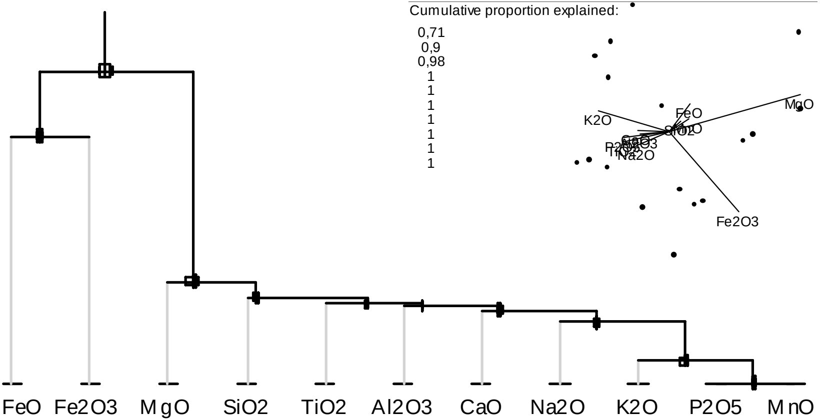


boxplot-scale: $(-4, +4)$

balance-dendrogram for Kilauea Iki data



balance-dendrogram and biplot for Kilauea Iki data



correlation matrix for balances of Kilauea Iki data

Coordinate Correlation Matrix

coord.num.	x2*	x3*	x4*	x5*	x6*	x7*	x8*	x9*	x10*
x1*	-0.800	0.735	0.724	-0.257	0.254	-0.005	0.159	-0.647	-0.676
x2*	1.000	-0.209	-0.176	-0.120	-0.271	-0.045	-0.324	0.355	0.169
x3*		1.000	0.981	-0.594	0.054	-0.058	-0.170	-0.628	-0.945
x4*			1.000	-0.598	0.180	0.038	-0.075	-0.695	-0.944
x5*				1.000	-0.125	-0.424	0.483	0.124	0.765
x6*					1.000	0.812	0.773	-0.683	-0.156
x7*						1.000	0.362	-0.297	-0.187
x8*							1.000	-0.645	0.220
x9*								1.000	0.562

scatterplot for balances of Kilauea Iki data