Application of compositional data analysis to geochemical data of marine sediments

H.H. Burger[1] and Th. Kuhn[2]
[1]Freie Universität Berlin, Berlin, Germany; *Heinz.Burger@FU-Berlin.de*
[2]ILeibnitz Institut für Polymerforschung Dresden, Germany; *kuhn@ipfdd.de*

Abstract

In an earlier investigation (Burger et al., 2000) five sediment cores near the Rodrigues Triple Junction in the Indian Ocean were studied applying classical statistical methods (fuzzy c-means clustering, linear mixing model, principal component analysis) for the extraction of endmembers and evaluating the spatial and temporal variation of geochemical signals. Three main factors of sedimentation were expected by the marine geologists: a volcano-genetic, a hydro-hydrothermal and an ultra-basic factor. The display of fuzzy membership values and/or factor scores versus depth provided consistent results for two factors only; the ultra-basic component could not be identified. The reason for this may be that only traditional statistical methods were applied, i.e. the untransformed components were used and the cosine-theta coefficient as similarity measure.

During the last decade considerable progress in compositional data analysis was made and many case studies were published using new tools for exploratory analysis of these data. Therefore it makes sense to check if the application of suitable data transformations, reduction of the D-part simplex to two or three factors and visual interpretation of the factor scores would lead to a revision of earlier results and to answers to open questions . In this paper we follow the lines of a paper of R. Tolosana-Delgado et al. (2005) starting with a problem-oriented interpretation of the biplot scattergram, extracting compositional factors, ilr-transformation of the components and visualization of the factor scores in a spatial context: The compositional factors will be plotted versus depth (time) of the core samples in order to facilitate the identification of the expected sources of the sedimentary process.

Kew words: compositional data analysis, biplot, deep sea sediments

1 Introduction

Deep sea sediments in the central valley and on the flanks of mid-ocean ridges represent different sedimentation processes. In general within the central valley volcanogenetic debris and pelagic sedimentation prevail. On the flanks of the rift axis settling of material from hydrothermally derived plumes becomes more important and may eventually overcome volcanogenetic debris. The transport of hydrothermally material to the flank sediments is realized via dispersion through the water column as plumes. The longer the plumes stay in the water column the more are they diluted until they reach sea water background (Figure 1).

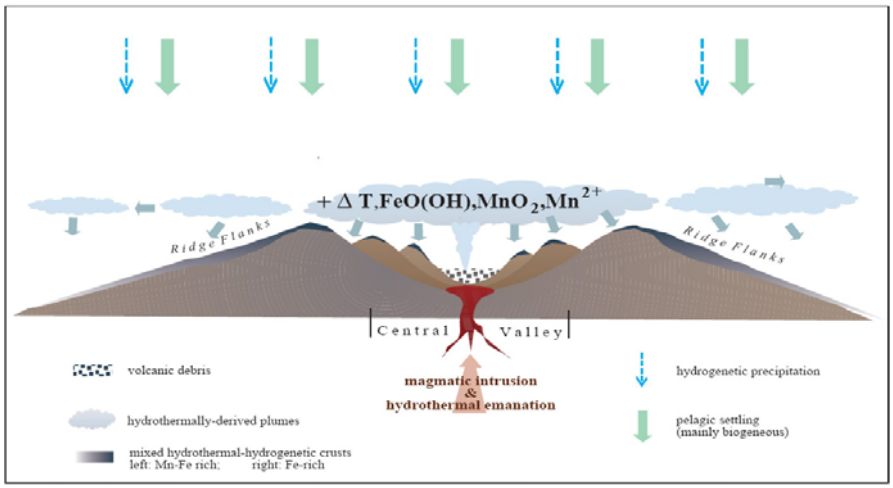


Figure 1: Schematic representation of the sedimentary processes on mid-ocean ridges

Following these processes surface sediments from different positions with respect to the rift axis should have different geochemical compositions. Analyzing a sediment core from one position, it may be possible to reconstruct the magmatic and hydrothermal history of the respective rift axis segment.

Five sediment cores from the flanks of the first Central Indian Ridge (CIR) segment and the first Southeast Indian Ridge segment (SEIR), seen from the Rodrigues Triple Junction (RTJ) were studied in order to evaluate the spatial and temporal variations of geochemical signals (Figure 2). These signals should reflect the dominating sedimentary environment and the related geological processes, i.e. the increasing distance from the mid-ocean ridge axis during ocean floor spreading as well as volcanic and hydrothermal events.

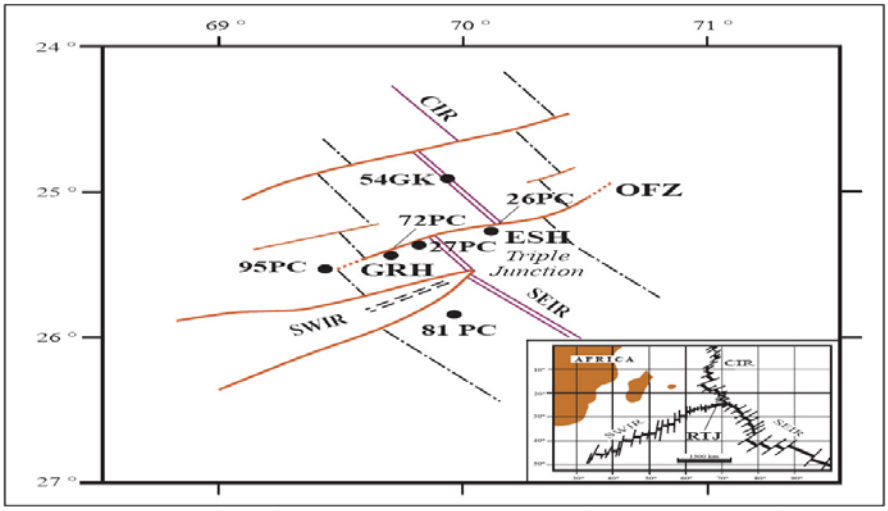


Figure 2: Location of sediment cores in the vicinity of the Rodrigues Triple Junction (RTJ). CIR – Central Indian Ridge, SEIR – Southeast Indian Ridge, SWIR – Southwest Indian Ridge, GRH – Green Rock Hill, OFZ – Oblique Fracture Zone, Numbers indicate location of piston cores.

The sediment cores were sampled at 10 cm intervals with respect to geochemical analysis (XRF) and in 2.5 cm intervals with respect to stratigraphic analysis (oxygene isotope stratigraphy). Three endmembers were expected by the marine geologists: volcanogenetic, hydrogenetic-hydrothermal and an ultrabasic factor. Fuzzy c-means cluster analysis (FCM) as well as algorithms for the extraction of extreme compositions were applied to model the endmembers and to estimate the contribution of each of them to the composition of the sediment samples.

In an earlier paper (Kuhn and others, 1996) fuzzy cluster analysis was applied in order to find natural groups within the dataset of all piston core samples. The original components were used with a cosine distance measure. The FCM-procedure produced two cluster centers with significant different geochemical composition which can be interpreted as volcanogenic and hydrogenetic-hydrothermal cluster. The ultrabasic factor could not be identified by this procedure. The result of this analysis is shown in Table 1.

Table 1: Element concentrations of centroids from fuzzy cluster analysis.

	Si (%)	Ti (%)	Al (%)	Fe (%)	Mg (%)	K (%)	Mn (%)	Cu (%)	Ni (%)	Zn (%)
Cc 1	**48,62**	**0,97**	13,99	22,91	9,99	**2,01**	1,37	0,075	0,033	0,029
Cc 2	21,17	0,87	13,63	**44,68**	13,72	1,69	**3,92**	**0,177**	0,073	0,068
Cc 3	31,22	0,85	13,72	37,34	11,48	1,96	3,17	0,13	0,075	0,050

The first centroid Cc1 is characterized by high contents of Si, Ti, Al and K. The second and third centroid show high values of Fe, Mg, Mn, Cu, Ni and Zn. There is no interpretable difference between the second and third centroid.

An algorithm of R. Renner (1996) was applied to extract end-members from these clusters. The first end-member represents volcanic detrital material (MORB) from the CIR and the second one is composed of

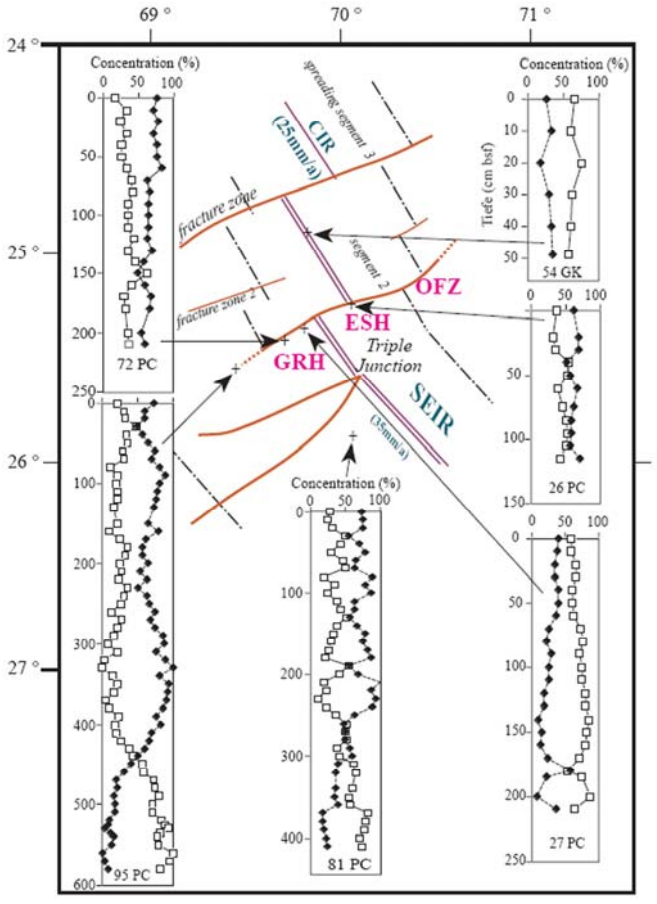


Figure 3: Vertical distribution of mixture proportions of two endmembers in the sediment cores versus depth below see floor. The squares represent the volcaogenetic endmember, the rhombs represent the hydrothermal-hydrogenetic endmember.

unspecified Fe-Mn-oxides with dominant Fe and minor clay content. The volcanic detritus dominates the lower parts of the sediment cores representing the Central Valley period whereas the autigenic-hydrothermal endmember dominates the flank periods. If we plot the contribution of each endmember in a sample versus depth we get a structure which shows very clearly the transitional change of the volcanogenetic endmember dominance to the hydrothermal-hydrogenetic dominance. This transition may mark the passing of the core locations from the central valley to the rift flanks (Figure 3).

2 Methodology

We first apply the biplot technique applied to clr-transformed data as proposed by Aitchison and Greenacre (2002).

$$clr(z_i) = \ln \frac{z_i}{\sqrt[D]{z_1 z_2 \ldots z_D}} \tag{1}$$

Our main interest focuses on the links between the variables, indication of collinearity of components and definition of subcompositions which should be related to the composition of the centroids and/or endmembers obtained earlier.

In a second step we choose an orthonormal basis, calculate the coordinates with respect to it and work on these coordinates (Pawlowsky-Glahn, 2003). A suitable basis was defined by Egozcue and others (2003) and the coordinates are obtained by

$$ilr(z_i) = \frac{1}{\sqrt{i(i+1)}} \ln \frac{z_1 z_2 \ldots z_i}{z_{i+1}^i}, \; i = 1,\ldots,D-1 \tag{2}$$

If we can extract two factors Φ_1, Φ_2 which contain the major part of the total variance of the data set in the coordinate space delivered by the biplot we can reduce the D-part simplex to three latent factors which form a composition in S^3 (for details see Tolosana and others, 2005).

$$\begin{pmatrix} clr(F_1) \\ clr(F_2) \\ clr(F_3) \end{pmatrix} = \begin{pmatrix} 1/\sqrt{6} & 1/\sqrt{2} \\ 1/\sqrt{6} & -1/\sqrt{2} \\ -2/\sqrt{6} & 0 \end{pmatrix} \begin{pmatrix} \Phi_2 \\ \Phi_1 \end{pmatrix} \tag{3}$$

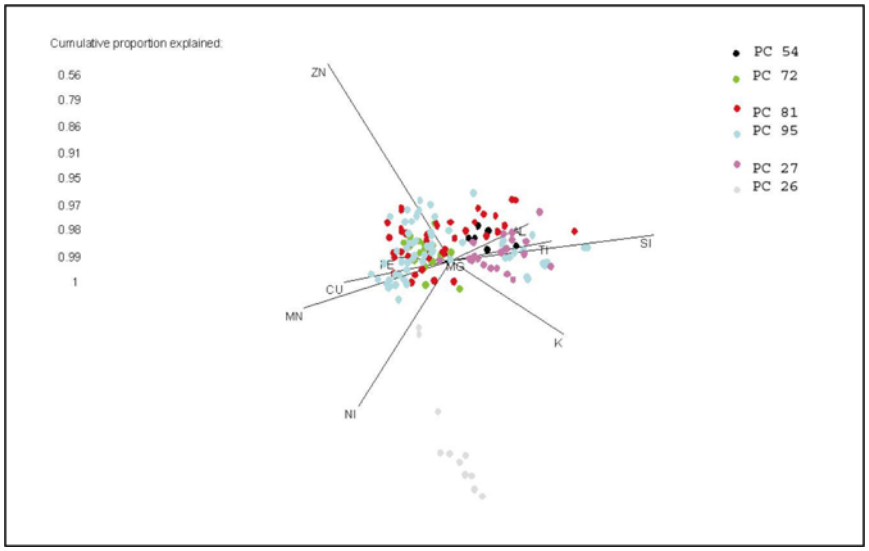
3 Exploratory data analysis

3.1 Analysis of biplots

We start the multivariate analysis of the deep sea piston core compositional data with the well known dimension reduction technique called biplot (details of this technique can be found e.g. in Aitchison and Greenacre, 2000). Similar to correspondence analysis the biplot shows the variables and the sample pattern in the same scatterplot which facilitates the interpretation of sample clusters in relation to variables or variable clusters.

Figure 4 shows the biplot of 10 components Si, Ti, Al, Fe, Mg, K, Mn, Cu, Ni, Zn (correctly: their oxides) and the factor scores of all piston samples (6 groups). Four principal components are required for representing about 90% of the total variance in the data set. It is obvious that PC 26 is located far outside the cluster of all other piston core samples so that the relation between the components may be dominated

by the contribution of this single group. Therefore PC 26 was eliminated from the multivariate analysis at the beginning and included again afterwards for final interpretation.



Figur**e 4:** Biplot of 10 components (vectors) and 165 samples from 6 deep sea cores.

The variation matrix of the complete dataset (Table 2) shows high clr-variances for Si, K, Ni, Zn and Mn which are related to the length of the vectors in the biplot. If we exclude PC 26 we get a significant reduction of the clr-variances of Zn and Ni which drop down to 0.1 and the values of E[ln(Cu/Ni)] and E[ln(Zn/Ni)] are close to zero which indicates that these components are approximately equal in the reduced dataset.

Table 2: Variation matrix for 10 components and the complete data set of 164 samples from 6 piston cores.

	SI	TI	AL	FE	MG	K	MN	CU	NI	ZN	clr Var
SI		0,14	0,16	0,42	0,27	0,23	0,77	0,58	0,64	0,77	0,27
TI	3,58		0,03	0,18	0,09	0,16	0,40	0,29	0,37	0,46	0,08
AL	0,83	-2,75		0,14	0,05	0,15	0,34	0,24	0,35	0,37	0,05
FE	-0,08	-3,66	-0,91		0,06	0,29	0,10	0,05	0,17	0,24	0,04
MG	1,00	-2,58	0,17	1,08		0,19	0,21	0,12	0,21	0,31	0,02
K	2,87	-0,70	2,04	2,95	1,87		0,47	0,36	0,38	0,67	0,16
MN	2,52	-1,06	1,68	2,59	1,52	-0,36		0,06	0,15	0,31	0,15
CU	5,58	2,00	4,75	5,66	4,58	2,71	3,06		0,14	0,24	0,08
NI	6,38	2,80	5,55	6,46	5,38	3,51	3,86	0,80		0,52	0,16
ZN	6,60	3,02	5,77	6,68	5,60	3,73	4,08	1,02	0,22		0,26
	Means								Tot	var	1,28

In the case of five groups the biplot shows that four principal components are necessary for representing about 90% of the total variance. The high clr-variance of K is dominating the graph of the biplot though according to experts potassium plays no known rôle for the sedimentation processes under consideration. It is a problem of multivariate statistical analysis that it always includes some arbitrariness in selecting or excluding variables and/or "outliers". This holds for classical statistical methods as well as for compositional data analysis. We were interested in finding some evidence for ultrabasic layers or components in the sediments which differ from basaltic rocks by higher contents in Ni and Mg. The high variance of K is regarded as noise factor and it's elimination reveals a biplot pattern which is easier to be interpreted in terms of the sedimentary processes described above (Figure 5): The components clr(Si), clr(Ti), clr(Al) located on the right hand side of the biplot are almost collinear while clr(Mn), clr(Cu), clr(Zn) build a cluster on the left hand side; the vector clr(Ni) is almost perpendicular to these components and the vector

of clr(Mg) is small (close to the origin of the biplot) which indicates that this component behaves like the geometric mean.

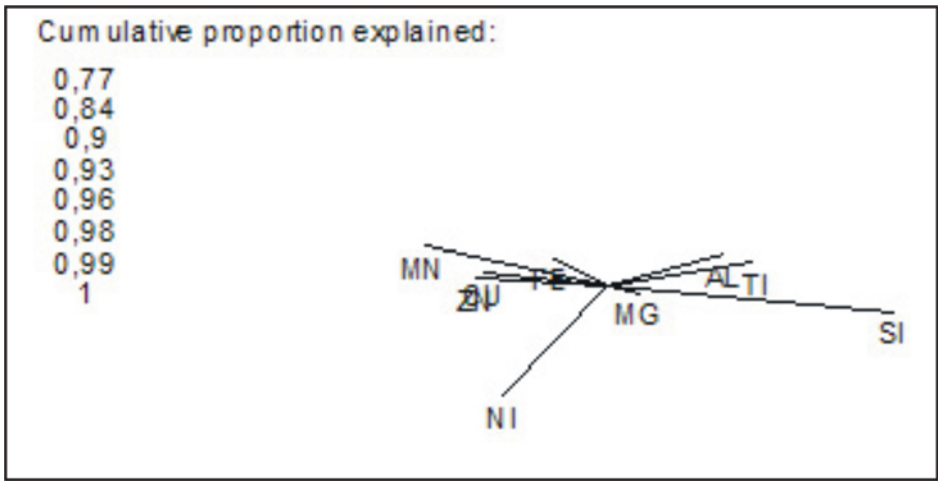


Figure 5: Biplot of 9 components from samples of 5 cores

This pattern of components in the biplot indicates that a more detailed analysis of subcompositions (Si, Ti, Al) and (Fe, Mn, Mg, Cu, Zn) and their relation to the component Ni may lead to the missing ultrabasic influence in the sediment samples.

3.2 Analysis of principal component scores (log contrasts)

We take the result from the visual inspection of the biplot as a starting point for the principal component analysis of two subcompositions i.e. we investigate the first eigenvectors of the covariance matrices of the clr-transformed data. In the case of (Si, Ti, Al) and (Fe, Mg, Mn, Cu, Zn) we get the associated logcontrasts

$$\Phi_1 = \frac{1}{\sqrt{122}}(9\ln(Si) - 4\ln(Ti) - 5\ln(Al))$$

$$\Phi_2 = \frac{1}{\sqrt{106}}(2\ln(Fe) - 5\ln(Mn) - 2\ln(Cu) - 3\ln(Zn))$$

A scattergram of the obtained logcontrasts are displayed in Figure 6. There exists a cluster of scores in the lower part of the scattergram which comprises samples from PC 72, PC 81 and PC 95 and on the right hand side three almost parallel branches which indicate high variability of Φ_2. The samples of PC 54 are obviously independent from Φ_1. When the samples are marked with the values of the variable cm_bsf (depth in cm below seafloor) no significant correlation correlation between this variable and the Φ-factor scores can be detected. It was expected by marine geologists that the composition of the sediment cores reflects the vertical evolution due to a changing environment like the river samples analyzed by Tolosana-Delgado and others (2005) who could clearly identify a downstream evolution of the Φ-factor scores.

The ternary diagram of the computed latent factors F_1, F_2, F_3 from Φ_1 and Φ_2 according to Equation (3) shows a similar pattern which contains no additional information. (Sample PC 81/330 cm_bsf is an outlier in both diagrams and is regarded as measurement error because values of samples just above and below are almost similar.)

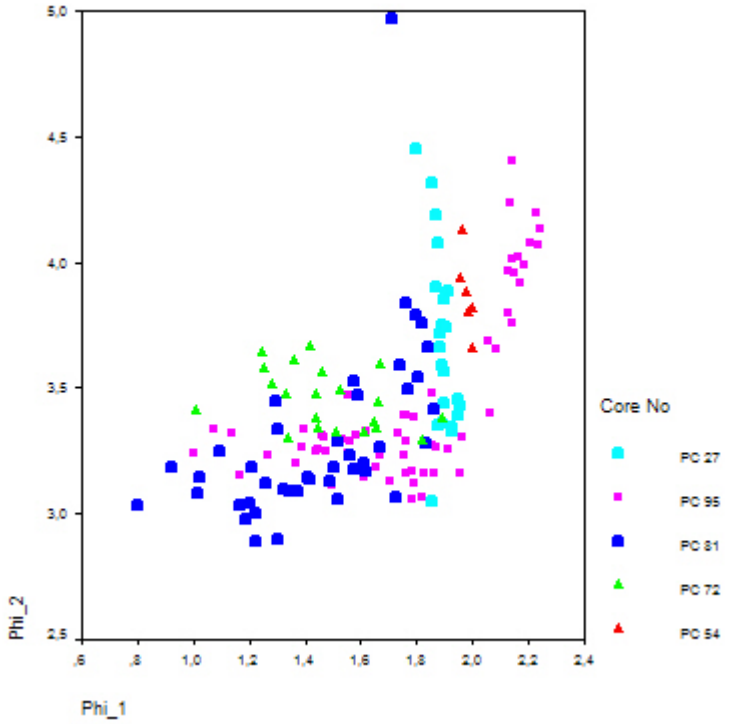


Figure 6: Scattergram of logcontrasts Φ_1 vs. Φ_2. Samples from different cores are characterized by different colors.

4 Clusteranalysis of compositions

Obviously there is no simple relationship between the logcontrasts and the evolution of the sediment in time but there is a well structured pattern in the scattergram of the logcontrasts which requires an explanation. Therefore we applied cluster analysis with Euklidian distance as similarity measure to the clr-transformed data to search for natural groups. Three centroids were obtained and the result of the back-transformed (clr > raw) components is shown in Table 3.

Table 3: Composition of cluster centroids back-transformed to percentage values

No	Si %	Ti %	Al %	Fe %	Mg %	K %	Mn %	Cu %	Ni %	Zn %
C1	**51,46**	**1,02**	**14,30**	20,50	9,78	1,84	0,96	0,05	0,02	0,021
C2	28,96	0,87	14,09	**38,70**	**12,25**	1,74	**3,11**	**0,14**	0,05	**0,057**
C3	32,56	0,87	11,62	35,99	11,86	**3,01**	3,75	0,14	**0,15**	0,008

A comparison with Table 1 shows that the compositional difference between the centroids obtained with raw data and suitable similarity measure (cosine) and clr-transformed data is small – except for Ni, Mn and K which have significant higher values now. (The advantage of working in the simplex is here that the centroids obtained are members of the simplex by definition.) Centroid C1 is characterized by high Si-Ti-Al-values which correspond to the first subcomposition selected in the preceding chapter. Centroid C2 with high value for Fe, Mg, Mn, Cu and Zn can be identified with the second subcomposition. Centroid C3 shows rather high concentrations of Mg, K and Ni which can be regarded as representative for the missing ultrabasic factor. C3 contains only samples from PC 26 which is located close to the OFZ (see Figure 2).

The following scatterplots in Figure 7 visualize the location of these three centroids and the two endmembers obtained earlier with respect to Φ_1, Φ_2 and Ni (which is not member of the subcompositions related to Φ_1 and Φ_2).

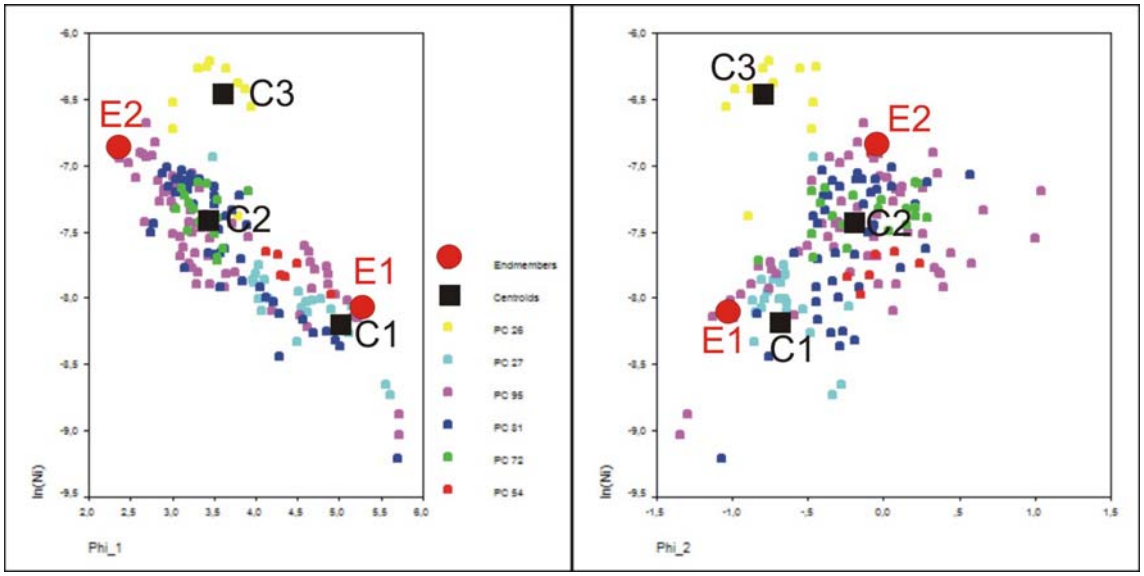


Figure 7: Scattergrams of factor Φ_1 and Φ_2 vs. ln(Ni). The centroids C1, C2, C3 are shown as black squares, the endmembers E1, E2 as red circles.

The endmembers E1, E2 are characterized by high/low values of Ni and Φ_1. Perpendicular to E1-E2 exists a considerable amount of variation which should be captured by one or two more endmembers. C1 is located in the neighborhood of E1 in both scattergrams in Figure 7 and can be interpreted as weathered volcanic detritus. C2 is located between E1 and E2 and may represent a transition from volcanic detritus to hydrothermal-hydrogenetic dominated sediments.

5 Conclusions

Deep sea piston core samples are documents of the tectonic and sedimentary evolution of the see floor. The analysis of these samples show that there is no simple evolutionary signal; the scattergram of various factor scores vs. time (here: depth below sea floor) reveals abrupt changes of the compositions, i.e. the membership of samples change within centimeters. These abrupt changes of cluster membership may be due to tectonic events, slumpings or hydrothermal plume fallout as it is commonly found along mid-ocean ridge flanks. The next step in this research will be a detailed analysis of the sediment composition from top to bottom which focuses on steady evolution (e.g. due to a changing environment when the oceanic crust is passing from the central valley to the rift flanks) in contrast to sudden changes (e.g. due to volcanic events).

Application of new methods in compositional data analysis may lead to a revision of earlier investigations when classical statistical methods were used. The bulk of new methods, transformations, operations and visualization tools require a problem-oriented approach, clearly defined questions and sound statistical methods. This can only be achieved by a close cooperation of experts in the geosciences and in statistics. The availability of software for compositional data analysis and the presentation of new answers for old problems will hopefully motivate the geoscientific community to apply these tools for their own problems.

Acknowledgements

We thank R. Tolosana-Delgado who spent a reasonable amount of time for valuable discussion on methodological and practical aspects of this work during a visit in Göttingen of one of us. Special thanks go to S. Thió-Henestrosa and J.A. Martín-Fernandez who develop and maintain the CoDaPack software package which was intensively used and which saved a lot of time and additional programming work.

References

Aitchison J. and M. Greenacre (2002): *Biplots for compositional data*.- Applied Statistics 51, 375 - 392

Bezdec, J. C., Ehrlich R. & W. Full (1984): *FCM: The fuzzy c-means clustering algorithm*.- Comp. & Geosci., 10, no.2, p 191-203.

Egozcue, J.J., V. Pawlowsi, G. Mateu-Figueras, and C. Barceló-Vidal (2003): *Isometric logratio transformations for compositional data analysis.* Mathematical Geology 35, 279 – 300.

Kuhn, T., Burger, H. & Halbach, P.: (2000): *Volcanic and hydrothermal evolution of ridge segments at the Rodrigues Triple Junction deduced from sediment geochemistry.* Marine Geology 169, p. 391-409.

Pawlowsky-Glahn, V. (2003): Statistical modeling on coordinates. See Thió-Henestrosa and Martin-Fernández (2003)

Renner, R. M. (1996): *An algorithm for constructing extreme compositions*.- Comp. & Geosci. 22, no. 1, pp 15-22.

Thió-Henestrosa and Martin-Fernández (Eds.) (2003): *Compositional Data Analysis Workshop – CoDaWork'03*, Proceedings, http://ima.udg.es/Activitats/CoDaWork03/. Universitat de Girona.

Tolosana-Delgado, R. N. Otero, V. Pawlowsky-Glahn and A. Soler (2005): *Latent compositional factors in the Llobregat River Basin (Spain) Hydrochemistry.* Mathematical Geology, 37, 683-706.