AN EXPERIMENTAL COMPARISON OF COKRIGING OF REGIONALIZED COMPOSITIONAL DATA USING FOUR DIFFERENT METHODS. CASE STUDY: BAUXITES IN HUNGARY

Eusebi Jarauta-Bragulat[a], **Carme Hervada-Sala**[b] , **Angela M. Diblasi**[c]

[a] eusebi.jarauta@upc.es Dept. de Matemàtica Aplicada III, ETSECCPB (UPC), Spain.
[b] carme.hervada@upc.es Dept. de Física i Enginyeria Nuclear, EUETIT (UPC), Spain.
[c] angelad@uncu.edu.ar Universidad Nacional de Cuyo, Argentina.
Website: ftp://ftp-urgell.upc.es/matematica/ejarauta/iamg2002/article1

ABSTRACT

An important problem in the geosciences is the estimation or prediction of regionalized compositions. In fact, it is usual to deal with data such as percentages, concentrations, ppm,...., and use them to estimate values in other locations. Compositional data have been regarded as difficult to work with because of the so-called constant sum constraint. Following Aitchison (1986), any meaningful statement about a composition can be expressed in terms of logratios, but those transformations, and their backtransformations, are not always easy to deal with. The aim of this paper is to compare results obtained applying different methodologies developed in geostatistics, with samples of compositional data from a bauxite deposit in Halimba II (Hungary). Firstly, a classical geostatistics study is done using raw data; secondly applying two wellknown transformations in compositional data analysis: additive logratio (ALR) and centered logratio (CLR); thirdly, the Fast Fourier Transform (FFT) methodology to calculate the spatial variance-covariance matrix is used in cokriging. To be able to compare predictive values and kriging errors respective backtransforms are found. At last, results obtained with the different approaches are discussed and compared.

1. Introduction

The most common goal in Geostatistics is to estimate the value of an unknown variable in a location using the information given by some samples in its surroundings. A new problem comes when compositional variables are studied, those variables are characterised by their constant sum; that is, variables summing up to one (proportions), summing up to 100 (percentages) and so on. Their main features have been studied and described by many authors (Aitchison, Barceló, Egozcue, Pawlowsky and others) and have settled some specific methodologies to work with them. Those methodologies set up some transformation of data; the best-known ones are average logratio (ALR) and centered logratio (CLR). Recently, Yao and Journel have found some way to calculate covariances matrix with Fast Fourier Transform (FFT). So, it seems sensible to apply all those methods, as well as the classical one used by most geologists to the same data to assess their applicability and results. This is the goal of this paper: the use of all those four methods, finding out their difficulties and comparing their results with some well-known data. The database is a set of compositional data from a bauxite deposit named Halimba, which is the largest one in Europe continuously mined since 1950. Gy.Bárdossy, Budapest, furnished the data.

2. The data set

The studied deposit is in Hungary (Europe) and it is limited by East 117.6 - 114.0; North 13.0 - 8.8 geographic coordinates in a topographic map. The deposit covers an area of more than 8 km^2; Halimba II is the only sector in the deposit that is still under prospection. The database consists of 55 samples representing 55 boreholes, after getting off 3 incomplete samples. In these boreholes the thickness of bauxite varies from 0.8 to 36.1 m. Variables used are the following: X = Easting; Y = Northing; V_1 = Concentration of Al_2O_3; V_2 = Concentration of

SiO_2; V_3 = Concentration of Fe_2O_3; V_4 = Concentration of TiO_2; V_5 = Concentration of H_2O; V_6 = Concentration of CaO; V_7 = Concentration of MgO; concentrations are in percent. The values of V_1 to V_7 represent weighted averages in each borehole taken from intervals of 0.5 to 1.0 m length. Full database and histograms of the variables can be found at our website; table 1 shows the descriptive statistics of data set.

Table 1. Descriptive statistics of data set.

	Range	Minimum	Maximum	Average	Standard deviation	Symmetry	Kurtosis
V1	8.3	49.9	58.2	54.569	2.234	- 0.647	- 0.558
V2	7.4	0.7	8.1	3.889	2.007	0.334	- 0.876
V3	7.4	20.4	27.8	23.698	1.898	0.523	- 0.096
V4	2.1	1.6	3.7	2.778	0.332	-0.683	2.773
V5	2.3	11.3	13.6	12.371	0.499	0.477	0.128
V6	2.7	0.1	2.8	0.536	0.545	2.232	5.543
V7	1.8	0.1	1.9	0.267	0.327	3.133	11.688

3. Raw data geostatistical analysis

This is a traditional method to estimate any regionalised variable in geostatistics. It consists of building up variograms for each variable and cross-variograms when there are more than one of them. Once experimental (cross)variograms have been built they must be modeled. The corresponding theoretical ones are used in (co)kriging system to estimate the values on a regular grid. Variograms for the seven variables have been calculated and modeled; a table with the full description of those models can be found in our website. Once all variograms were built, cokriging has been done using KB2D program from GSLIB (1998).

4. Geostatistical analysis considering variables as compositions: ALR transform

Classical applications of geostatistics are related to mapping the spatial distribution of the variables under study. They give emphasis to characterize the variogram model and use the kriging (error) variance as a measure of estimation accuracy. Nowadays, some problems have been reported with compositional data. Those problems have been studied by many authors (references [1], [2], [3] and [8]). The main problem when handling compositional data is the so-called constant sum (K) constraint. Usually $K = 1$ or $K = 100$, if data are percentages. So, if $V_1, ..., V_N$ are proportions of N elements, then $V_1 + \cdot \cdot \cdot + V_N = K$, which means that variables are not independent. To deal with compositional data and avoid this constraint, Aitchison has proposed some transforms. We have used two of them: average logratio (ALR) and centered logratio (CLR). With those transformations variables become independent and then classical kriging can be performed. As it is said beforehand, $V_1, ..., V_N$ must follow the constant sum constraint, but this quite never is true. Actually, we must define a new variable (called the residual) as $V_R = K - (V_1 + \cdot \cdot \cdot + V_N)$. Then, the ALR variables U_i (i = 1,2,…, N) are defined as follows:

$$U_i = alr(V_i) = \log \frac{V_i}{V_R}, i = 1,2,...,N \tag{1}$$

So now we are working with several U_i variables which do not follow the sum constraint; so they can be used as any other geostatistical data. We build and model their variograms (they can be found in our website). Once variogram was built, kriging has been done using KB2D program from GSLIB (1998). Kriging results must be backtranformed to have the estimation of V_i in the grid; in this case, the corresponding ALR-backtranform is:

$$V_i = \frac{\exp U_i}{1+\sum_{i=1}^{N} \exp U_i} K, i = 1,2,\ldots,N \qquad (2)$$

5. Geostatistical analysis considering variables as compositions: CLR transform

Once V_R has been defined, we define a new variable as the geometrical mean of all of them:

$$V_{gm} = (V_1 V_2 \ldots V_N V_R)^{\frac{1}{N+1}} = \exp\left(\frac{1}{N+1} \left(\sum_{i=1}^{N} (\ln V_i + \ln V_R \right) \right) \qquad (3)$$

Then, CLR transform consists in stating N+1 new variables as:

$$W_j = clr(V_i) = \ln \frac{V_j}{V_{gm}}, j = 1,2,\ldots, N+1 \qquad (4)$$

These N+1 variables are not constrained, so they can be modeled and estimated. Once variograms have been built (they can be found in our website) cokriging has been done using KB2D program from GSLIB (1998). Then, backtransforms must be done to recover original variables. CLR-backtranform is:

$$V_j = \frac{\exp W_j}{\sum_{i=1}^{N+1} \exp W_i}, \quad j = 1,2,\ldots, N+1 \qquad (5)$$

6. Fast Fourier Transform method to calculate the covariance matrix

On the other hand, to avoid the modeling of variograms and crossvariograms, which may be very subjective, Yao and Journel (1998) have developed the so-called FFT method, which can be applied, in principle, to any kind of data. With FFT you do not need the independence of the variables and it builds up the covariance matrix, which can be used directly to krige. This approach works as follows:

a) Generate an experimental correlogram map on a regular grid. The grid typically has multiple nodes without estimates. The user has to specify the minimum number of data to be considered in the estimates at every node. This task is performed by program CORRMAP (see reference [6]).

b) Program INTMAP fills in the blanks typically present in the grid generated in step 1 by using a smooth local interpolation.

c) Program MULTSMTH corrects the smoothed grid to generate a third grid that is a tabulation of a positively semidefined correlogram. This condition is required to assure a unique solution for the kriging system of equations yielding a non-negative kriging variance.

d) Convert the correlogram tabulation in step 3 to covariance tabulation by multiplying the correlogram grid by the sampling variance.

e) As it was not possible to use KB2D to krige, because with this method we obtain the covariance matrix and not the variograms, we had to change it (see reference [5]).

7. Results and discussion

Table 2 shows descriptive statistics for the estimations. Variables shown are raw estimations (V_i), backtranformations of ALR estimations (BACK U_i), backtransformations of CLR estimations (BACK W_i) and estimations using FFT (FFT V_i). Table 3 shows descriptive statistics of their differences, that is the differences between raw estimations and each of the other estimations. Kriging errors can be found in website. Figures comparing kriging results for the seven variables can be found in website; as an example you can see hereafter, in figure 1, results for variable V_1. In this figure, (a) is refered to raw data, (b) to the back-transformation of ALR-variable, (c) to the backtransformation of CLR-variable and (d) to the FFT transformation method. Looking at the contour maps, no significant differences among the first three methods arise. However (d)-picture, the one belonging to FFT method, shows higher resolution. It seems that it is because this method is less subjective.

8. Conclusions.

Using the results of this study some conclusions can be built:

a) Kriging results in Halimba II using the four methods are quite similar.

b) As regarding to the kriging errors, comparison is not so easy because it is not true that the backtransform of ALR and CLR transformations belong to the same space as the data (this is why Martin et al. defined stress).

c) FFT method seems to be the best one, because it is less subjective, more precise and, furthermore, it is the easiest method to use. However, this method does not take into account if data are compositional or not.

9. References.

[1] Aitchison, J. (1997). *The one-hour course in compositional data analysis or compositional data analysis simple.* Proceedings of IAMG'97. IMNE, Barcelona. Part I. pp.3-35.

[2] Aitchison, J. *The Statistical Analysis of Compositional Data*. Chapman and Hall, 1986.

[3] Barceló, C. et al. *Mathematical Foundations of Compositional Data Analysis*. The 2001 annual conference of IAMG, Cancún (Mexico), 20 pp.

[4] Deutsch, C.V. and A.G. Journel (1998). *Geostatistical software library and user's guide – GSLIB.* Oxford University Press, 1 CD + 369 pp.

[5] Hervada-Sala, C. and E. Jarauta-Bragulat (2001). *Modifications to kb2d program in GSLIB to allow use of tabulated covariances calculated with Fast Fourier Transform method.* Computers & Geosciences, vol.27, num. 07, pags 887-889.

[6] Ma, X and Yao,T. (2001). *A program for 2D modeling (cross)correlogram tables using Fast Fourier Transform.* Computers & Geosciences, vol.27, num. 07, pags 763-774.

[7] Martín-Fernández, J.A. et al. (2001). *Criteria to compare estimation methods of regionalized compositions.* Mathematical Geology, vol 33, num. 8, pags 889-909.

[8] Olea, R.A. (1999). *Geostatistics for engineers and earth scientists.* Kluwer Academic Publishers, 303 pp.

[9] Pawlowsky, V. et al. (1995). *Estimation of regionalized compositions: a comparison of three methods.* Mathematical Geology, 27(1), 105–127.

[10] Yao, T. and A.G. Journel (1998). *Automatic modelling of (cross)covariance tables using fast Fourier transform.* Mathematical Geology, 30(6), 589–615.

Acknowledgements.

Authors want to thank specially:

- To Department of Applied Mathematics III, UPC.
- To the School of Civil Engineering (ETSECCPB) in UPC.
- To the "Consorci Escola Industrial de Barcelona" (CEIB).

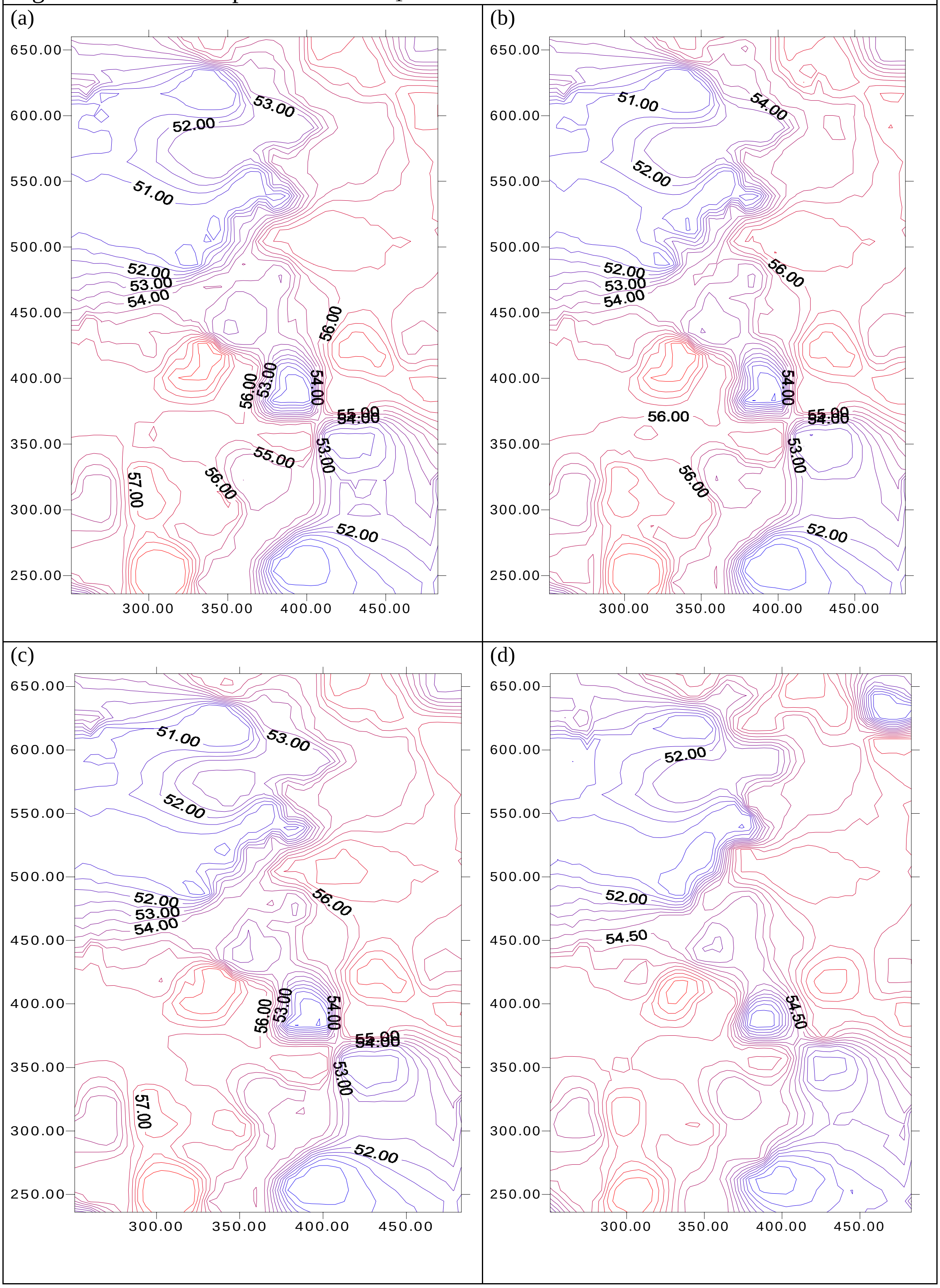
Table 2. Descriptive statistics of kriging estimations.

Variable	**N**	**Average**	**Median**	**Std. Dev.**	**Minimum**	**Maximum**
V_1	845	54.62	55.20	2.16	49.90	58.20
BACK U_1	845	54.68	55.19	2.14	49.78	58.10
BACK W_1	845	54.66	55.18	2.14	49.88	58.17
FFT V_1	578	54.62	55.30	2.24	49.90	58.20
V_2	845	3.929	3.800	2.000	0.700	8.100
BACK U_2	845	3.898	3.626	2.007	0.700	8.150
BACK W_2	845	3.899	3.623	2.006	0.700	8.150
FFT V_2	578	3.924	3.602	2.011	0.700	8.100
V_3	845	23.70	23.45	1.72	20.40	27.80
BACK U_3	845	23.72	23.41	1.72	20.32	27.75
BACK W_3	845	23.73	23.51	1.72	20.46	27.89
FFT V_3	578	23.61	23.30	1.83	20.40	27.80
V_4	845	2.777	2.800	0.348	1.600	3.700
BACK U_4	845	2.779	2.814	0.348	1.600	3.710
BACK W_4	845	2.779	2.813	0.346	1.600	3.680
FFT V_4	578	2.783	2.801	0.341	1.600	3.700
V_5	845	12.34	12.30	0.45	11.30	13.60
BACK U_5	845	12.35	12.34	0.47	11.01	13.76
BACK W_5	845	12.35	12.26	0.45	11.38	13.62
FFT V_5	578	12.37	12.30	0.50	11.30	13.60
V_6	845	0.504	0.300	0.465	0.100	2.800
BACK U_6	845	0.490	0.301	0.453	0.090	2.810
BACK W_6	845	0.490	0.301	0.453	0.100	2.810
FFT V_6	578	0.503	0.300	0.489	0.100	2.800
V_7	845	0.247	0.100	0.264	0.100	1.900
BACK U_7	845	0.234	0.105	0.247	0.090	1.900
BACK W_7	845	0.234	0.101	0.247	0.100	1.900
FFT V_7	578	0.249	0.100	0.293	0.100	1.900

Table 3. Descriptive statistics of estimation differences.

	N	**Average**	**Median**	**Std. Dev.**	**Minimum**	**Maximum**
V_1 - BACK U_1	845	-0.0621	-0.0088	0.2640	-2.6162	0.6762
V_1 - BACK W_1	845	-0.0429	0.0009	0.2104	-1.7219	0.3092
V_1 - FFT V_1	560	0.0189	0.0000	0.6570	-3.1083	6.7002
V_2 - BACK U_2	845	0.0310	-0.0001	0.2165	-1.2449	2.0673
V_2 - BACK W_2	845	0.0301	-0.0011	0.1993	-1.1678	1.7569
V_2 - FFT V_2	560	0.0026	0.0000	0.4389	-3.9004	2.8890
V_3 - BACK U_3	845	-0.0148	-0.0121	0.1170	-0.9911	0.4930
V_3 - BACK W_3	845	-0.0299	-0.0262	0.1234	-1.1340	0.7786
V_3 - FFT V_3	560	0.0174	0.0000	0.6029	-3.5003	6.5000
V_4 - BACK U_4	845	-0.0018	0.0023	0.0268	-0.2413	0.2638
V_4 - BACK W_4	845	-0.0017	0.0005	0.0260	-0.1852	0.2491
V_4 - FFT V_4	560	0.0047	0.0000	0.0667	-0.6930	0.4000
V_5 - BACK U_5	845	-0.0044	0.0034	0.1598	-1.0208	1.2711
V_5 - BACK W_6	845	-0.0072	-0.0039	0.0685	-0.6018	0.2735
V_5 - FFT V_6	560	-0.0116	0.0000	0.1633	-1.0505	1.0498
V_6 - BACK U_6	845	0.0138	-0.0001	0.0633	-0.0745	0.6584
V_6 - BACK W_6	845	0.0142	0.0000	0.0681	-0.0285	0.7951
V_6 - FFT V_6	560	-0.0054	0.0000	0.1530	-1.2472	1.2500
V_7 - BACK U_7	845	0.0133	0.0001	0.0673	-0.0251	0.6202
V_7 - BACK W_7	845	0.0135	0.0001	0.0666	-0.0057	0.5916
V_7 - FFT V_7	560	-0.0040	0.0000	0.1144	-1.0331	0.9000

Figure 1. Contour maps of variable V_1.


(a)
650.00
600.00
550.00
500.00
450.00
400.00
350.00
300.00
250.00
300.00
350.00
400.00
450.00
53.00
52.00
51.00
52.00
53.00
54.00
56.00
56.00
53.00
54.00
55.00
54.00
55.00
53.00
57.00
56.00
52.00
(b)
650.00
600.00
550.00
500.00
450.00
400.00
350.00
300.00
250.00
300.00
350.00
400.00
450.00
51.00
54.00
52.00
52.00
53.00
54.00
56.00
54.00
56.00
55.00
54.00
53.00
56.00
52.00
(c)
650.00
600.00
550.00
500.00
450.00
400.00
350.00
300.00
250.00
300.00
350.00
400.00
450.00
51.00
53.00
52.00
52.00
53.00
54.00
56.00
56.00
53.00
54.00
55.00
54.00
53.00
57.00
52.00
(d)
650.00
600.00
550.00
500.00
450.00
400.00
350.00
300.00
250.00
300.00
350.00
400.00
450.00
52.00
52.00
54.50
54.50