COMPARISON OF KRIGING RESULTS OF REGIONALISED COMPOSITIONAL DATA USING THREE DIFFERENT DATA TRANSFORMATIONS. CASE STUDY: BAUXITES IN HUNGARY

Eusebi Jarauta-Bragulat[1], Carme Hervada-Sala[2]

[1] Dept. Applied Mathematics III, ETSECCPB/EUETIB,
Universitat Politècnica de Catalunya (UPC), Spain. Eusebi.Jarauta@upc.es
[2] Dept. Physics and Nuclear Engineering. EUETIT,
Universitat Politècnica de Catalunya (UPC), Spain. Carme.Hervada@upc.es

ABSTRACT

In geosciences it is usual to deal with regionalized data such as percentages, concentrations, mg/kg (ppm), that is, compositional data; well-known problems have been pointed out with these kind of data, specially related to geostatistical analysis and kriging, because of the so-called constant sum constraint. From Aitchison's research about compositional data (1986), any meaningful statement about a composition has to be expressed in terms of data transformation; several data transformations have been proposed to deal with compositional data: additive logratio (alr) and centered logratio (clr) are the most used transformations until now. Recently Egozcue and Pawlowsky have proposed a new data transformation with the aim to have an orthonormal basis in compositional space (simplex); this transformation is called isometric logratio (ilr).

In this paper, we study these data transformations and the application of alr, clr and ilr transformations to a data set which consists in samples of compositional data from a bauxite deposit in Halimba II (Hungary) and we compare obtained results by kriging with those different transformations and applying the Fast Fourier Transform (FFT) methodology to calculate the spatial variance-covariance matrix. A previous work with the same data set can be found in Jarauta-Bragulat and others (2002).

Introduction

In Geosciences, a common goal is the estimation of unknown variable values in some points, usually a regular grid, from information given by samples in its surroundings. A particular and interesting case is 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. The main features of this compositional variables 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: the additive logratio transformation (ALR), the centered logratio transformation (CLR), and recently, the isometric logratio transformation (ILR), defined by Egozcue and others (2003).

It seems sensible to apply all those methods and compare obtained results, in order to determine which one runs better to analyse data information and to obtain a better estimation in a grid. This is the goal of this paper: the use of all three transformations, finding out their difficulties and comparing their results with some well-known data; the Fast Fourier Transform (FFT) is applied to calculate the covariances matrix. 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. G. Bárdossy, Budapest, furnished the data.

The data set

The studied deposit is in Hungary (Europe) and its limits are 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 described in Table 1(a); in all cases, concentrations are expressed in percentages. The values of V_1 to V_7 represent weighted averages in a borehole taken from intervals of 0.5 to 1.0 meters length. Full database and histograms of the variables can be found at our website. Table 1(b) shows the descriptive statistics of data and Figure 1 illustrates the scatterplot of data location.

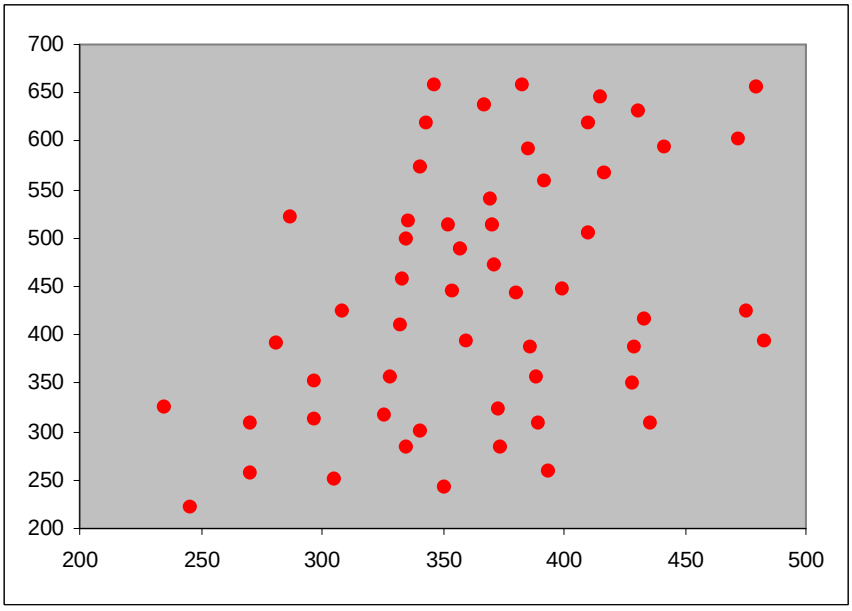
Table 1 (a). Description of used variables.

Variable	Notation
Eating	X
Northing	Y
Concentration of Al_2O_3	V1
Concentration of SiO_2	V2
Concentration of Fe_2O_3	V3
Concentration of TiO_2	V4
Concentration of H_2O	V5
Concentration of CaO	V6
Concentration of MgO	V7

Table 1(b). Descriptive statistics of data set.

	V1	V2	V3	V4	V5	V6	V7
Mean	54.57	3.89	23.70	2.78	12.37	0.54	0.27
Standard error	0.30	0.27	0.26	0.04	0.07	0.07	0.04
Median	55.20	3.60	23.30	2.80	12.30	0.30	0.10
Mode	56.00	1.70	27.30	2.90	12.50	0.20	0.10
Standard deviation	2.23	2.01	1.90	0.33	0.50	0.55	0.33
Variance	4.99	4.03	3.60	0.11	0.25	0.30	0.11
Kurtosis	-0.56	-0.88	-0.10	2.77	0.13	5.54	11.69
Symmetry	-0.65	0.33	0.52	-0.68	0.48	2.23	3.13
Range	8.30	7.40	7.40	2.10	2.30	2.70	1.80
Minimum	49.90	0.70	20.40	1.60	11.30	0.10	0.10
Maximum	58.20	8.10	27.80	3.70	13.60	2.80	1.90

Figure 1. Scatterplot of sample data locations.



Compositional variables and transformations.

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 (Aitchison, J. (1997), Aitchison, J. (1986), Jarauta-Bragulat and others (2002), Pawlowsky and others (1995)). The main problem when handling compositional data is the so-called constant sum (K_{CL}) constraint; usually we have $K_{CL}=1$ if data are parts per unit, $K_{CL}=100$ if data are percentages, and $K_{CL}=10^6$ if data are parts per milion (mg/kg). It is known that this sum constraint means that variables are not independent. Usually, the data information consists in a matrix of *n* columns (variables) and *m* rows (cases); so, if $V_1, V_2, ..., V_n$ are the compositional studied variables, then we define the residual variable as follows:

$$V_R = V_{n+1} = K_{CL} - (V_1 + V_2 + \cdots + V_n)$$

and now the data matrix has $n+1$ columns. In any case, if $\sum_{j=1}^{n+1} x_j = K_{CL}$, then $\sum_{j=1}^{n+1}\left(\frac{x_j}{K_{CL}}\right)=1$. This is the reason why, from hereafter, we take as unit the constant sum K_{CL}.

The set of compositional data, called the (*n*+1)-dimensional simplex, is denoted by $\mathcal{S}_{n+1}$; it is a real vector space with the inner sum called *perturbation* and the external product called *power transformation*. To deal with compositional data and avoid the constant sum constraint, some transformations have been proposed:

a) Additive logratio transformation.
Given any element $x=(x_1,x_2,\ldots,x_n,x_{n+1})\in\mathcal{S}_{n+1}$ of the simplex, we consider the map $f:\mathcal{S}_{n+1}\to\mathbb{R}^n$ defined as follows:

$$f(x)=\left(\ln\frac{x_1}{x_{n+1}},\ln\frac{x_2}{x_{n+1}},\ldots,\ln\frac{x_n}{x_{n+1}}\right);\ x=(x_1,x_2,\ldots,x_{n+1})\in\mathcal{S}_{n+1} \qquad (3.1)$$

This map is an isomorphism of vector spaces, it is called *additive logratio transforma-tion*, and it is denoted by ALR.

b) Centered logratio transformation.
Given any element $x=(x_1,x_2,\ldots,x_n,x_{n+1})\in\mathcal{S}_{n+1}$ of the simplex, we define the map $\varphi:\mathcal{S}_{n+1}\to\mathbb{R}^{n+1}$ as following:

$$\varphi(x)=\left(\ln\frac{x_1}{g(x)},\ln\frac{x_2}{g(x)},\ ...,\ln\frac{x_{n+1}}{g(x)}\right);\ g(x)=g(x_1,...,x_n)=\sqrt[n+1]{x_1x_2\cdots x_{n+1}} \qquad (3.2)$$

This map is called *centered logratio transformation*, and it is denoted by CLR. Is easy to see that the image set $H=\varphi(\mathcal{S}_{n+1})$ is the hyperplan of $\mathbb{R}^{n+1}$ given by the equation $y_1+y_2+\cdots+y_{n+1}=0$.

c) Isometric logratio transformation.
Given any element $x=(x_1,x_2,\ldots,x_n,x_{n+1})\in\mathcal{S}_{n+1}$ of the simplex, we consider the map $\psi:S_{n+1}\to H$ defined by:

$$\psi(x)=\left(\sqrt{\frac{1}{2}}\ln\frac{x_1}{x_2},\sqrt{\frac{2}{3}}\ln\frac{\sqrt{x_1.x_2}}{x_3},...,\sqrt{\frac{k}{k+1}}\ln\frac{g(x_1,...,x_k)}{x_{k+1}},\ ...,\sqrt{\frac{n}{n+1}}\ln\frac{g(x_1,...,x_n)}{x_{n+1}}\right) \tag{3.3}$$

where $g(x_1,...,x_k)$ indicates the geometric mean, as before. This map is called *isometric logratio transformation*, and it is denoted by ILR.

Compositional variables and back-transformations.

Once compositional data have been transformed according to one of the above expressions and kriging has been done, kriging results must be backtranformed to have the estimation of data values on a, usually, regular grid. So, we need the expressions of corresponding back-transformations as follows.

a) Additive logratio back-transformation.

If $(y_1,y_2,...,y_n)$ are the ALR coordinates of an element $x=(x_1,x_2,\ldots,x_n,x_{n+1})$ of the símplex, then back-transformations must be obtained by applying the following expression:

$$x_i=\frac{\exp(y_i)}{1+\sum_{i=1}^{n}\exp(y_i)}\quad (i=1,2,...,n);\qquad x_{n+1}=1-(x_1+\cdots x_n) \tag{4.1}$$

Values must be multiplied by the constant K_{CL} when it is different than the unit.

b) Centered logratio back-transformation.

If $(z_1,z_2,...,z_{n+1})$ are the CLR coordinates of an element $x=(x_1,x_2,\ldots,x_n,x_{n+1})$ of the símplex, then back-transformations must be obtained by applying the following expression:

$$x_j=\frac{\exp(z_j)}{\sum_{i=1}^{n+1}\exp(z_j)};\ \ j=1,2,...,n+1. \tag{4.2}$$

c) Isometric logratio back-transformation.

If $(u_1,u_2,...,u_n)$ are the ILR coordinates of an element $x=(x_1,x_2,\ldots,x_n,x_{n+1})$ of the símplex, then back-transformations must be obtained by applying the following expression:

$$x_k=\frac{\exp(S_k)}{1+\sum_{k=1}^{n}\exp(S_k)}\quad (k=1,2,...,n);\qquad x_{n+1}=1-(x_1+\cdots+x_n) \tag{4.3}$$

where:

$$\begin{aligned}&\exp\left(\frac{1}{2}\sqrt{2}u_1+\frac{1}{3}\sqrt{\frac{3}{2}}u_2+\cdots+\frac{1}{j+1}\sqrt{\frac{j+1}{j}}u_j+\cdots+\frac{1}{n}\sqrt{\frac{n}{n-1}}u_{n-1}+\sqrt{\frac{n+1}{n}}u_n\right)=\exp(S_1)\\&\exp\left(-\frac{k-1}{k}\sqrt{\frac{k}{k-1}}u_{k-1}+\frac{1}{k+1}\sqrt{\frac{k+1}{k}}u_k+\cdots+\frac{1}{n}\sqrt{\frac{n}{n-1}}u_{n-1}+\sqrt{\frac{n+1}{n}}u_n\right)=\exp(S_k)\\&(k=2,3,...,n)\end{aligned} \tag{4.4}$$

Results and discussion

Transformations described in the before paragraph have been used on our data set; with these data transformed values we want to krige them on the same grid and compare obtained results. To avoid the modeling of variograms and crossvariograms, which may be very subjective, Yao and Journel (1998) have developed the so-called FFT method. This method can be applied 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. To do all this data processing, the following setps have been done:

1) Find the normal score of all data transformations. This task is performed with NSCORE program of GSLIB.
2) 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.
3) Program INTMAP fills in the blanks typically present in the grid generated in step 2 by using a smooth local interpolation.
4) 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.
5) Convert the correlogram tabulation in step 4 to covariance tabulation by multiplying the correlogram grid by the sampling variance.
6) As it was not possible to use KB2D to krige, because with this method we obtain the covariance matrix and not the variograms, we apply the program developed by Hervada-Sala and others (2001).
7) Obtain the ALR, CLR and ILR kriged values with back normal score transformation.
8) Back transformation of ALR, CLR and ILR kriged values to return to the simplex initial space.

Table 2 shows descriptive statistics for the ALR, CLR and ILR back-transformation kriging values estimations. And table 3 shows the average kriging errors using all the three transformations.

Conclusions.

All estimations have been done using the same method: kriging on correlogram matrices calculated by FFT methodology, so this work can be used to compare the three transformations. The main conclusions of this work can be stated as follows:

1. The transformation that fits the best the results seems to be ALR because its statistics are more alike the raw data statistics.
2. The reason why ALR fits best is that residual values where very low. If residuals values were high, we expect ILR would fit the best.
3. ALR is also the easiest of the transformations to be carried out.

References.

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.

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

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

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.

Jarauta-Bragulat, E., C. Hervada-Sala and Angela M. Diblasi (2002). *An experimental comparison of cokriging of regionalized compositional data using four different methods. Case study: bauxites in Hungary*. IAMG annual confference, Berlin.

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.

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

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: the Department of Applied Mathematics III (UPC), the School of Civil Engineering (UPC) and the "Consorci Escola Industrial de Barcelona" (CEIB), for its support in this contribution.

Table 2. Descriptive statistics of back transformation kriging values (on each row, first value is back-alr, second is back-clr and third is back-ilr).

	V1	V2	V3	V4	V5	V6	V7
Mean	54,60	3,91	23,65	2,77	12,36	0,51	0,25
	52,78	3,80	22,93	2,68	11,94	0,49	0,24
	54,51	3,94	23,73	2,78	12,35	0,52	0,25
Standard error	0,09	0,09	0,08	0,02	0,02	0,02	0,01
	0,10	0,08	0,08	0,01	0,02	0,02	0,01
	0,11	0,09	0,08	0,02	0,02	0,02	0,01
Median	55,19	3,80	23,35	2,82	12,35	0,30	0,10
	53,34	3,69	22,72	2,73	11,87	0,29	0,10
	55,17	3,79	23,34	2,85	12,23	0,30	0,10
Mode	49,78	7,75	27,32	1,60	12,15	0,20	0,10
	48,50	7,47	26,62	1,55	11,72	0,19	0,10
	49,92	7,72	27,33	1,59	12,04	0,20	0,10
Standard deviation	2,21	2,07	1,83	0,36	0,50	0,51	0,30
	2,46	1,98	1,89	0,35	0,43	0,49	0,28
	2,62	2,05	1,90	0,37	0,53	0,53	0,30
Variance	4,86	4,27	3,35	0,13	0,25	0,26	0,09
	6,05	3,93	3,56	0,12	0,19	0,24	0,08
	6,84	4,22	3,62	0,14	0,28	0,28	0,09
Kurtosis	-0,53	-0,93	-0,04	2,06	-0,06	4,77	10,53
	-0,13	-0,92	-0,05	1,73	0,13	4,91	10,99
	36,71	-0,94	1,96	5,67	11,16	5,59	11,98
Symmetry	-0,64	0,33	0,53	-0,68	0,43	2,14	2,99
	-0,70	0,31	0,41	-0,59	0,61	2,18	3,07
	-3,49	0,30	0,84	-0,11	1,75	2,29	3,18
Range	8,32	7,45	7,41	2,12	2,35	2,71	1,80
	12,72	8,14	9,43	2,04	2,44	2,57	1,69
	35,39	7,42	14,80	3,87	5,96	3,01	2,00
Minimum	49,78	0,70	20,33	1,60	11,28	0,10	0,10
	44,18	0,68	19,44	1,55	11,04	0,10	0,10
	22,80	0,70	20,42	1,51	11,21	0,10	0,10
Maximum	58,10	8,15	27,73	3,71	13,62	2,81	1,90
	56,90	8,82	28,88	3,59	13,48	2,67	1,78
	58,19	8,12	35,23	5,38	17,17	3,11	2,10

Table 3. Average error estimates by kriging corresponding to ALR, CLR and ILR transformations.

V1	V2	V3	V4	V5	V6	V7
-0,94649	-0,93020	-0,94915	-0,95002	-0,94978	-0,91356	-0,93039
1,08166	1,05306	1,18499	1,09415	1,11746	0,95566	1,18951
1,05425	1,14255	1,01301	1,07581	0,98225	1,24854	1,07890