An affine equivariant anamorphosis for compositional data

K. G. VAN DEN BOOGAART[1] , R. TOLOSANA-DELGADO and U. MUELLER[2*]

[1]Helmholtz Institute for Resources Technology Freiberg, Germany
[2]School of Engineering, Edith Cowan University, Australia u.mueller@ecu.edu.au
[*] presenting author

Abstract

For the geostatistical treatment of compositional data it s common to trnsform the data to logratios. Several logratio transformations are available and invariance of the results under the choice of logratio transform is desirable, but this is not automatically satisfied for geostatistical simulation where it is common that the data are first mapped to Gaussian space. The usual method, the normal score transform, is not independent of the choice of logratio nor are the transformed data multivariate normal. In this contribution a method is proposed based on an affine-equivariant kernel density estimation, which is then continuously deformed to a multivariate standard normal distribution. The anamorphosis is achieved via the co-deformation of the underlying space. The method is illustrated and compared with existing alternatives using a case study from a Westaustralian iron ore mining operation.

1 Introduction

Compositional data analysis proposes a simple workflow for the (geo)statistical treatment of data in percentages, ppm, proportions, etc, data which represent the relative mass/weight/importance of several components forming a system. This is: transform the data with an appropriate logratio transformation, analyse the transformed scores, and finally back-transform the obtained results (model coefficients, predictions, interpolations, simulations, etc). Several logratio transformations (alr, clr, ilr) have been proposed, and in some applications one might have advantages over other. However, for geostatistical applications, most often the question is to obtain equal (or equivalent) results with any of the transformations. Cokriging has been proven to yield interpolations invariant with respect to the choice of logratio (Tolosana-Delgado, 2006), a property known as *affine equivariance*.

In the case of simulation algorithms this invariance also holds, as long as the assumption of Gaussianity necessary for simulation is honored. However, for most data sets it is necessary to map data to a Gaussian space prior to simulation. The most commonly used transformation is the normal score transform where a transformation to normal scores is achieved via quantile matching. Another method achieves Gaussian anamorphosis via an expansion of the cdf as a Hermite series (Rivoirard, 1984), but even though in the multivariate case variables are transformed jointly to normal scores, the resulting transformed data are not multivariate normal, even though they have standard normal marginal distributions. More recently two methods have been presented to transform a multivariate data set into a multivariate normal data set. The first of these is the stepwise conditional transformation (Leuangthong and Deutsch, 2003). This method is hierarchical, with the first variable being transformed to normal scores based on a quantile transformation, the second variable being transformed to normality conditional on the first and so on. The ordering of the variables is recommended to be based on continuity of the input data. The structure of the input data is restored during the backtransformation, and no distributional assumptions about the input data are made. The transformation results in data that are independent, but there is no guarantee that they are spatially independent. The projection pursuit method (Barnett et. al., 2014) is based on the PPDE method of Friedman and Tukey (1974), and as a first step the individual variables are transformed to normal scores via a quantile matching, they are then centered and sphered. Following this the PPDE algorithm is applied iteratively to yield multivariate normal data. One of the steps in each iteration

is a normal score transform by means of a quantile matching. As was the case for the stepwise conditional transformation the transformed variables are uncorrelated at lag 0 by construction, but this may not be the case for non-zero lag separations. What is common in all of these methods is the reliance on a quantile matching to obtain normal scores, and as a result a lack of affine equivariance. That is, the transformations depend on the choice of logratio and so the use of one or another (log)ratio transform will lead to different sets of normal scores which in general do not have any relationship between them. Thus conditional simulations will depend on the (arbitrary) choice of logratio transform, an undesirable characteristic. It is therefore necessary to develop an alternative anamorphosis that has the equivariance property. In this paper one such method is introduced and its features are explored.

2 Compositional transformations

A data set is considered of compositional nature if its variables describe the relative importance of some components forming a whole. Typically, this relative importance is described in % or other proportional units (parts in one, ppm, ppb, ppt, and so on). Geochemical data sets are archetypical examples of compositional data, formed by non-negative variables (as a negative proportion of one element is impossible), and for each datum, all variables should sum up to 100% or less.

In a composition $\mathbf{y} = [y_1, y_2, \ldots, y_D]$ with D components, these conditions are stated as $y_i \geq 0$ and $\sum_{i=1}^{D} y_i \leq 100\%$. When identified with a point in D-dimensional real space $\mathbb{R}^D$, the set of points satisfying these conditions is called the D-part simplex, denoted $\mathcal{S}^D$.

To account for the relative nature of the information inherent in the data, ratios of components should be analysed, instead of the raw components and for mathematical reasons, it is better to analyse logratios than ratios (Aitchison, 1986). Several families of logratios have been proposed as standard tools. They include the pairwise logratio transformation (plr, (Aitchison, 1986)), the additive logratio transformation (alr, (Aitchison, 1986)),the centred logratio transformation (clr, (Aitchison, 1986)) and the isometric logratio transformation (ilr, (Egozcue et. al. , 2003)). A pairwise logratio transformation expresses a composition with the $D(D-1)$ possible pairwise logratios $\zeta_{ij} = \ln(y_i/y_j)$). For the alr transformation, a family of logratios is computed relative to a fixed component, often the last: $\zeta_i = \ln(y_i/y_D)$. In matrix notation this can be written as

$$\mathrm{alr}(\mathbf{y}) = \mathbf{F} \cdot \ln \mathbf{y}, \quad \mathbf{F} = [\mathbf{I}_{D-1}, -\mathbf{1}_{D-1}],$$

For the centred logratio transformation D scores are computed as logarithms of quotients of the components and the geometric mean of the components $g(\mathbf{y}) = \sqrt[D]{\prod_j y_j}$: $\zeta_i = \ln(y_i/g(\mathbf{y}))$. In matrix notation the transformation can be written as

$$\mathrm{clr}(\mathbf{y}) = \mathbf{H} \cdot \ln \mathbf{y}, \quad \mathbf{H} = \mathbf{I}_D - \frac{1}{D}\mathbf{1}_D \cdot \mathbf{1}_D^t,$$

where the logarithm is applied component-wise, the matrix $\mathbf{I}_D$ is the $D \times D$ identity matrix and $\mathbf{1}_D$ is a (column-)vector with D ones. The inverse of the clr transformation is

$$\mathrm{clr}^{-1}(\boldsymbol{\zeta}) =: \mathcal{C}[\exp(\boldsymbol{\zeta})] = \mathbf{y}, \tag{1}$$

where the closure operation

$$\mathcal{C}[\mathbf{y}] = \frac{100}{\mathbf{1}_D^t \cdot \mathbf{y}}\mathbf{y}$$

forces the argument to a constant sum 100% without changing the ratios between the components.

Finally, an isometric logratio transformations can be defined by means of a $(D-1) \times D$ matrix $\mathbf{V}$

$$\mathrm{ilr}(\mathbf{y}) := \mathbf{V} \cdot \ln \mathbf{y},$$

where $\mathbf{V} \cdot \mathbf{V}^t = \mathbf{I}_{D-1}$ and $\mathbf{V} \cdot \mathbf{1}_D = \mathbf{0}_{D-1}$, i.e. which columns are orthogonal vectors and each sums up to zero. These logratio transformations are related through the equivalences

$$\text{ilr}(\mathbf{y}) = \mathbf{V} \cdot \text{clr}(\mathbf{y}), \qquad \text{clr}(\mathbf{y}) = \mathbf{V}^t \cdot \text{ilr}(\mathbf{y}) \tag{2}$$

$$\text{alr}(\mathbf{y}) = \mathbf{F} \cdot \text{clr}(\mathbf{y}), \qquad \text{clr}(\mathbf{y}) = \mathbf{H}_* \cdot \text{alr}(\mathbf{y}), \tag{3}$$

where $\mathbf{H}_*$ represents the matrix $\mathbf{H}$ without the last column. With these expressions and Eq. (1), the inverse isometric logratio transformation is given as $\text{ilr}^{-1}(\boldsymbol{\zeta}_V) = \mathcal{C}[\exp(\mathbf{V}^t \cdot \boldsymbol{\zeta}_V)]$, where $\boldsymbol{\zeta}_V$ denote the ilr-transform of a composition relative to $\mathbf{V}$.

To statistically describe a random compositional data set $\mathbf{Y}$, expected value and variability can be conmputed using the logratio transformed data. The choice of logratio transformation is irrelevant, as the statistics in each representation can be recovered from any other representation. With regard to the means,

$$\text{clr}^{-1}(\text{E}[\text{clr}(\mathbf{Y})]) = \text{ilr}^{-1}(\text{E}[\text{ilr}(\mathbf{Y})]) = \text{alr}^{-1}(\text{E}[\text{alr}(\mathbf{Y})]) = \text{cen}(\mathbf{Z}), \tag{4}$$

the closed geometric mean, by definition called the geometric center of the random composition (Aitchison, 1986). To characterize variability, one can use the covariance matrices of alr-, clr- respectively ilr-transformed compositions, $\boldsymbol{\Gamma} = [\gamma_{ij}] = \text{Var}[\text{alr}(\mathbf{Y})]$, $\boldsymbol{\Psi} = [\psi_{ij}] = \text{Var}[\text{clr}(\mathbf{Y})]$ and $\boldsymbol{\Sigma}_V = \text{Var}[\text{ilr}(\mathbf{Y})]$. These matrices are all related to each other through the linear relations (Aitchison, 1986; Egozcue et. al. , 2003)

$$\boldsymbol{\Sigma}_V = \mathbf{V} \cdot \boldsymbol{\Psi} \cdot \mathbf{V}^t, \qquad \boldsymbol{\Psi} = \mathbf{V}^t \cdot \boldsymbol{\Sigma}_V \cdot \mathbf{V}, \tag{5}$$

$$\boldsymbol{\Gamma} = \mathbf{F} \cdot \boldsymbol{\Psi} \cdot \mathbf{F}^t, \qquad \boldsymbol{\Psi} = \mathbf{H}_*^t \cdot \boldsymbol{\Gamma} \cdot \mathbf{H}_*, \tag{6}$$

All these expressions (4-6) hold for the population expectation and variance of a theoretical random composition, because the mean vector and the covariance matrix are affine equivariant properties of the random composition. Moreover, for any compositional sample, most commonly used estimators of these quantities are also affine equivariant, and thus satify the same equivalences (Eqs. 4-6). This is the case for the classical estimators (moment estimators and maximum likelihood estimators, Tolosana-Delgado, 2006) as well as some robust estimators (minimum covariance determinant covariance and mean, Filzmoser and Hron, 2008), but not for rank-based estimators (median, interquartile range), as the data are ordered in different non-compatible ways with each logratio.

3 Some concepts of multivariate geostatistics

Geostatistics offer a set of tools for modeling the spatial dependence of a data set, with the focus on obtaining estimates or simulations of its variables, based on the formalism of random functions (Matheron, 1965). Consider a domain $\mathcal{D}$ in 2D or 3D space. Let x be the coordinates of any point within that domain. A vector-valued random function is the collection of all random vectors indexed by this spatial index, $\mathbf{Z}(x)$. The vector-valued random function $\phi(\mathbf{Z}(x))$ is multivariate gaussian if it is fully determined by its mean value $\boldsymbol{\mu}(x) = \text{E}[\phi(\mathbf{Z}(x))]$ and its covariance function $\mathbf{C}(x, x') = \text{var}[\phi(\mathbf{Z}(x)), \phi(\mathbf{Z}(x'))]$,

We further assume that the property of *intrinsic stationarity* is satisfied: the expectation and variance of the relevant *increments* of the random function are not spatially-dependent, $\text{E}[\mathbf{Z}(x+h) - \mathbf{Z}(x)] = 0$ and $\text{var}[\mathbf{Z}(x+h) - \mathbf{Z}(x)] = 2\boldsymbol{\Gamma}(h)$. The function $\boldsymbol{\Gamma}(h)$, the well-known *(semi)-variogram*, depends only on the lag displacement h between the sample locations ($h = x' - x$), but not on the exact locations.

Once a model of the variogram is available, standard geostatistical techniques for example, *ordinary cokriging* or *cosimulation* to obtain values for the ϕ-transformed vector at unsampled locations, $\phi(\mathbf{Z}_0) = \phi(\mathbf{Z}(x_0))$, which can be back-transformed to obtain predictions or simulations in the original scale.

In the case of a compositional random function, the mean $\boldsymbol{\mu}$ is a vector of the same length as $(\mathbf{Z})$ (i.e. D if the clr is used, and $D-1$ if the alr is used), and its alr and clr versions are related through Eq. (3). In the same way, variograms and covariance functions are matrices of $D \times D$ or $(D-1) \times (D-1)$ elements (respectively with clr and alr transformations), and the two versions of these matrices are linked to each other through Eq. (6).

4 Flow characterization of a multivariate anamorphosis

The multivariate anamorphosis proposed here is based on ideas borrowed from Lagrangian mechanics. We will construct a flow that shifts the data towards the center of the multivariate normal distribution. The anamorphosis is achieved via the co-deformation of the underlying space. Given a set $\{\mathbf{z}_i, i = 1, \ldots, n\} \in \mathbb{R}^D$ of vectors with variance-covariance matrix Σ, assume that each vector $\mathbf{z}_i$ represents the center of a smoothing kernel K_i. From a Lagrangian perspective the motion of the kernel can be described by the location of its center at time t

$$\mathbf{z}_i(t) = (1-t)\mathbf{z}_i$$

and the time dependent spread of the smoothing kernel is assumed to be a linear function of time:

$$\sigma(t) = (\sigma_1 - \sigma_0)t + \sigma_0 \tag{7}$$

Thus at time $t = 0$ the center of the kernel $K_i(0)$ is located at $\mathbf{z}_i$ and its spread is equal to σ_0 and at time 1, the kernel $K_i(1)$ has center $\mathbf{0}$ and spread σ_1. Time dependent normal scores are defined relative to the variance covariance matrix of the data and the spread of the kernel at time t by

$$\mathbf{s}_i(\mathbf{z}, t) = \mathbf{L}^{-1}\frac{\mathbf{z} - \mathbf{z}_i(t)}{\sigma(t)} \tag{8}$$

where $\mathbf{z}$ denotes a point of the kernel and the matrix $\mathbf{L}$ is the lower triangular factor in the Cholesky decomposition of the variance covariance matrix: $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^t$. The local movement for the mass of the smoothing kernel K_i is given by:

$$X(\tilde{t}; \mathbf{z}, i, t) = \mathbf{z}_i(\tilde{t}) + \mathbf{L}\mathbf{s}_i(\mathbf{z}, t)\sigma(\tilde{t}) \tag{9}$$

and the corresponding speed is

$$\begin{aligned} \mathbf{v}_i(\mathbf{z}, t) &= \frac{\partial}{\partial \tilde{t}} X(\tilde{t}; \mathbf{z}, i, t) \\ &= -\mathbf{z}_i + \frac{\sigma_1 - \sigma_0}{\sigma(t)}(\mathbf{z} - \mathbf{z}_i(t)) \end{aligned}$$

Next define the mean speed as

$$\mathbf{v}(\mathbf{z}, t) = \frac{\sum w_i \mathbf{v}_i(\mathbf{z}, t)}{\sum w_i}$$

where the weights at time t are defined as

$$w_i(\mathbf{z}, t) = \exp\left(-\frac{1}{2}\|\mathbf{s}_i(\mathbf{z}, t)\|^2\right),\ i = 1, \ldots, n$$

The equation of motion from the raw data to Gaussian space then is given by

$$\frac{\partial}{\partial t}\mathbf{g}(t) = \mathbf{v}(\mathbf{g}(t), t)$$

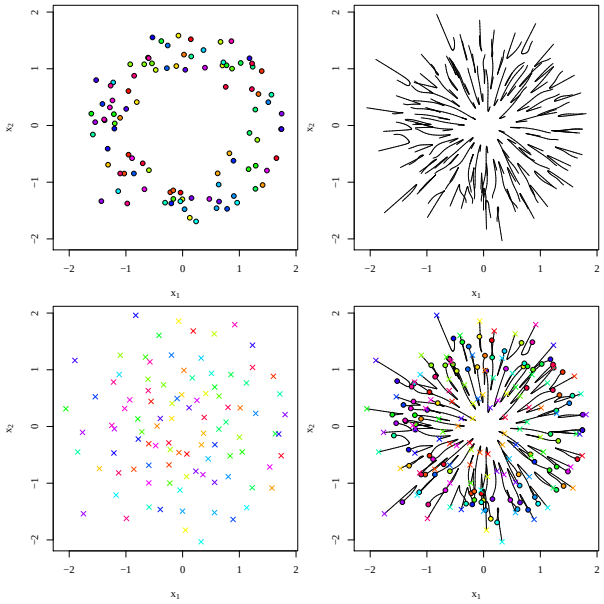

Figure 1: Trajectories for the flow transformation

where $\mathbf{g}(t)$ denotes the position of a point at time t and that from Gaussian space back to raw data space is:

$$\frac{\partial}{\partial t}\mathbf{r}(t) = -\mathbf{v}(\mathbf{r}(t), 1-t)$$

These equations of motion will be solved numerically to obtain the normal scores and back-transforms respectively.
The above flow transformation has the following properties:

- The flow is invariant under affine transformations.
  Suppose that the points and smoothing kernel centers are subjected to the affine transformation

  $$\zeta = \mathbf{A}\mathbf{z} + \mathbf{b}$$

  and that the movement at points at time t is given as $\zeta(t) = \mathbf{A}\mathbf{z}(t) + (1-t)\mathbf{b}$, when the affine transformation is taken into account, then the motion field at the transformed locations is

  $$\begin{aligned}
  \mathbf{v}_i(\zeta, t) &= -\zeta_i + \frac{\sigma_1 - \sigma_0}{\sigma(t)}(\zeta(t) - \zeta_i(t)) \\
  &= -\mathbf{A}\mathbf{z}_i - \mathbf{b} + \frac{\sigma_1 - \sigma_0}{\sigma(t)}(\mathbf{A}\mathbf{z}(t) + (1-t)\mathbf{b} - (\mathbf{A}\mathbf{z}_i(t) + (1-t)\mathbf{b})) \\
  &= -\mathbf{A}\mathbf{z}_i - \mathbf{b} + \frac{\sigma_1 - \sigma_0}{\sigma(t)}(\mathbf{A}\mathbf{z}(t) - (\mathbf{A}\mathbf{z}_i(t))) \\
  &= \mathbf{A}(-\mathbf{z}_i + \frac{\sigma_1 - \sigma_0}{\sigma(t)}(\mathbf{z}(t) - \mathbf{z}_i(t))) - \mathbf{b} \\
  &= \mathbf{A}\mathbf{v}_i(\mathbf{z}, t) - \mathbf{b}
  \end{aligned}$$

- The normal scores depend on the initial spread of the kernels σ_0.
  This is a direct consequence of the scaling impacting on the time dependent normal scores and so resulting in slightly different speeds of the flow field.

An example of the flow anamorphosis is shown in Figure 1. Here the initial bivariate distribution of the data is annular (top left). The application of the anamorphosis proceeds by first moving points towards the origin and then pushing them back out towards the periphery (bottom right), resulting in a more circular scatter plot (bottom left). In the process, neighbourhood relationships are preserved: data pairs that were close originally are close in the final bivariate distribution and vice-versa.

5 Case Study

5.1 Data description

This study used a single bench of 6 m long blast hole (BH) samples from an iron ore mine located in the central Yilgarn, Western Australia (Ward and Mueller, 2012). The bench consists of five discrete rotated fault imbricates. The longest strike distance within each horst (constrained by the angular tolerance) is 60 meters. only the western part of the bench is considered. Seven analytes measured in weight percent were available within the dataset, although only the three main elements of interest for iron ore mining (Fe, SiO_2 and Al_2O_3) were examined. As these analytes on their own form a subcomposition, a filler variable r was introduced in order to satisfy the constant sum constraint. The data exhibit the typical hematite enriched iron ore distributions; negatively skewed Fe, positively skewed SiO_2 and Al_2O_3 distributions (Table 1). The scatterplots also display typical behavior; strong negative correlation both between Fe and Al_2O_3, between Fe and SiO_2 (-0.91 and -0.90), and strong positive correlation between SiO_2 and Al_2O_3 (0.93).

Table 1: Descriptive statistics of BH analytes

Variable	n	Min	Max	Mean	Std. Dev.	CoV
Al_2O_3	400	0.12	8.53	1.42	1.43	1.01
Fe	400	51.09	69.45	63.94	2.76	0.04
SiO_2	400	0.34	10.09	2.22	1.81	0.82
r	400	27.77	35.34	32.10	1.18	0.04

The data were alr-transformed by putting

$$alrX = ln(X/r) \tag{10}$$

where $X \in \{Al_2O_3, Fe, SiO_2\}$. Of the three variables, only $alrFe$ follows a normal distribution (Table), and the correlations between the transformed variables are -0.35, -0.35 and 0.93 for $Al_2O_3 - Fe$, $SiO_2 - Fe$ and $SiO_2 - Al_2O_3$ respectively. A transformation to normal scores is therefore required prior to any simulation of the data.

Table 2: Descriptive statistics of alr-variables

Variable	n	Min	Max	Mean	Std. Dev.	SW
$alrAl_2O_3$	400	-5.61	-1.23	-3.54	0.91	7.051e-05
$alrFe$	400	0.48	0.88	0.68	0.05	0.9943
$alrSiO_2$	400	-4.61	-1.06	-2.96	0.74	0.9717

5.2 Flow anamorphosis and projection pursuit transform

To apply the flow anamorphosis, the setting of the initial density σ_0 needs to be determined. The following criteria were used to determine a suitable value: the multivariate normality of the transformed data, symmetry and spread (as measured by the interquatile range) of the distributions. In the case of the data considered here, the resulting multivariate distribution is multivariate normal for values $\sigma_0 \leq 0.5$ and the resulting univariate distributions are 'reasonably symmetric for $\sigma_0 \leq 0.05$. It was therefore decided to set $\sigma_0 = 0.05$. In addition the three transformed variables are uncorrelated (correlation coefficients for $V1-V2$, $V1-V3$ and $V2-V3$ are -0.005, 0.000 and 0.003 respectively) and therefore independent. Consideration of the omnidirectional experimental semivariograms and cross-variograms (see Figure 2) further shows that the transformed variables are spatially uncorrelated (with a mean value of Tercan's τ (Tercan, 1999) equal to 0.043).

Table 3: Descriptive statistics of FA-variables and PPMT variables

Variable	n	Min	Max	Mean	Std. Dev.
$FlowAnaV1$	400	-2.42	2.87	-0.0007	0.947
$FlowAnaV2$	400	-2.48	2.36	0.0111	0.931
$FlowAnaV3$	400	-2.42	2.38	-0.0054	0.929
$ppmtV1$	400	-2.91	3.21	0	0.999
$ppmtV2$	400	-2.77	2.88	0	0.998
$ppmtV3$	400	-2.83	2.97	0	0.998

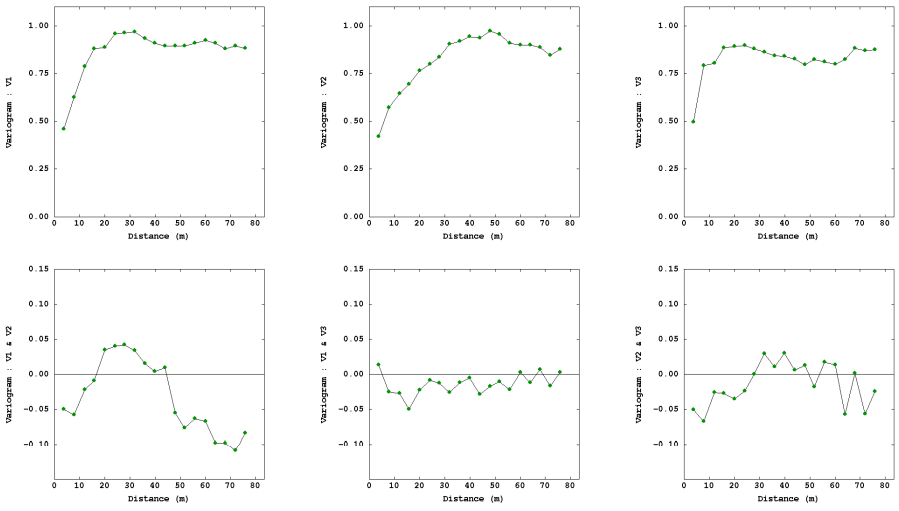

Figure 2: Omnidirectional experimental semivariograms and cross variograms for FA-transformed data

Like the flow normal scores the ppmt normal scores result in variables that follow a multivariate normal distribution. It should be noted that the standard deviations for the transformed variables are closer to unity than in the case of the flow-anamorphosis, a consequence of the repeated application of the nscore transform in ppmt. The three transformed variables are uncorrelated (correlations between the variables are 0.0011, -0.0022 and 0.0034) and just as in the case of the flow-anamorphosis the omnidirectional experimental semivariograms and cross-variograms show little evidence of spatial correlation (Tercan's τ equal to 0.045).

Univariate simulation can therefore be used to generate realisations of the FA and PPMT normal scores. The three variables show anisotropy which is consistent with the strike direction of the western part of the bench, nested structures consisting of a nugget and up to 3 spherical transition structures were required to achieve the mix of zonal and geometric anisotropies in the data. Similarly, parameters for the direct variograms of the ppmt-variables indicate a mix of zonal and geometric anisotropy.

The models were validated using leave one out cross-validation with simple kriging as the

Table 4: Parameters for variogram models of FA and PPMT variables

Variable	Nugget	Direction	Sill	Range	Sill	Range	Sill	Range
$FlowAnaV1$	0.30	N70	0.59	(26.5, 29)	0.17	$(\infty, 17)$		
$FlowAnaV2$	0.31	N80	0.15	(8, 8)	0.43	(48, 40)	0.17	$(\infty, 48)$
$FlowAnaV3$	0.12	N80	0.51	(8.2, 8.2)	0.25	(30, 8.2)		
$ppmtV1$	0.31	N80	0.68	(22, 22)	0.14	$(\infty, 22)$		
$ppmtV2$	0.50	N80	0.55	(49, 35)	0.14	$(\infty, 35)$		
$ppmtV3$	0.25	N80	0.50	(9, 9)	0.12	(33, 10)		

estimation algorithm. In each case a minimum of 4 samples and a maximum of 12 samples were used within an ellipse of major axis 30m and minor axis 20m oriented according to the direction of greatest continuity of the relevant variogram model. The results in Table 5 demonstrate that the models are suitable for simulation.

Table 5: Cross-validation results for variogram models of FA and PPMT variables

Variable	$\bar{e}_{st}$	Var(e_{st})	corr(Z,Z^*)	Corr($Z*$,e_{st})
$FlowAnaV1$	-0.007	1.021	0.626	0.009
$FlowAnaV2$	-0.003	0.879	0.666	-0.038
$FlowAnaV3$	-0.020	0.909	0.559	-0.026
$ppmtV1$	-0.006	1.095	0.603	0.043
$ppmtV2$	-0.002	0.838	0.667	-0.043
$ppmtV3$	-0.021	0.922	0.541	-0.028

5.3 Simulation results

Turning bands simulation with 500 bands was used to generate 100 realisations of each of the 6 variables, followed by back transformation first to alr-space and then back to the simplex. The simulations based on the flow-anamorphosis have reproduced the sample statistics reasonably well: For Fe, the sample mean is reproduced, while for Al_2O_3 and SiO_2 there is slight underestimation of the target, standard deviations tend to be slightly underestimates with the sample standard deviations falling between the upper quartile and the maximum of the realisation statistic. The overall reproduction of the Cdf is satisfactory (see Figure 3) and the correlations between the three variables are close to those of the input data. The shape of the scatter diagrams also approximate those of the input data (see Figure 4). An inspection of the spatial maps of the realisations showed no artefacts. For the projection pursuit method, the statistics are still acceptable, but not as well reproduced, which is also evidenced in the qq-plots of realisations against samples, notably overestimation in low values for Fe and underestimation of high values for high values of the remaining attributes (see Figure 3).

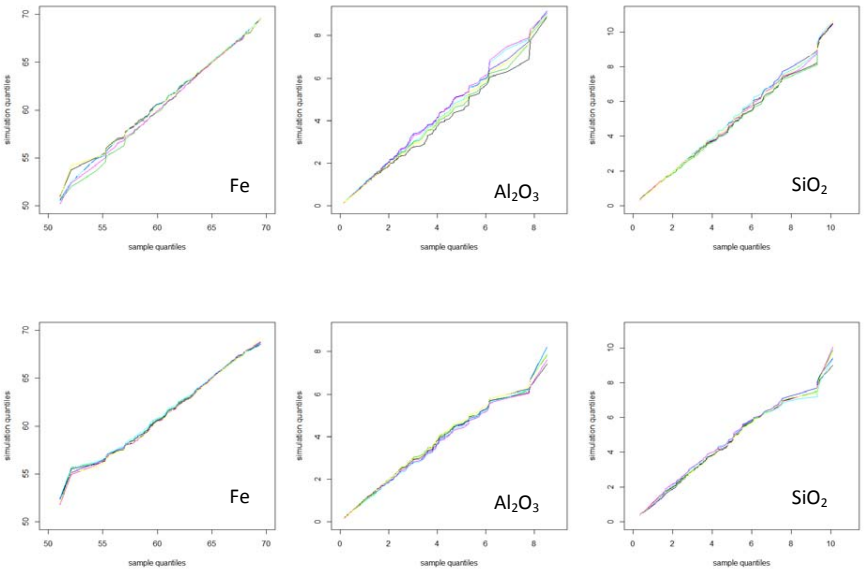

Figure 3: QQplots of selected realisations against input data : FA realisations (top) , PPMT realisations (bottom)

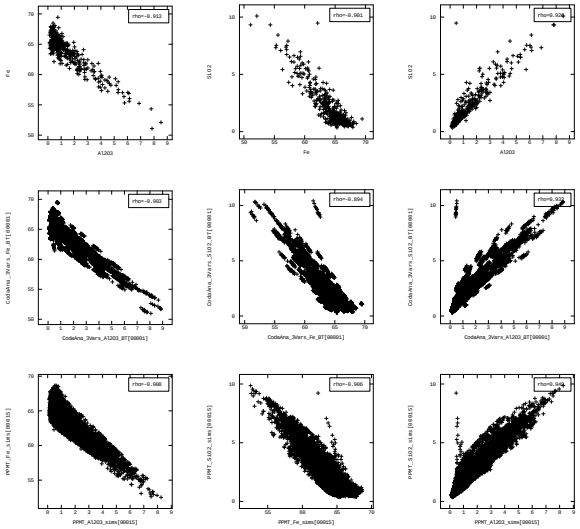

Figure 4: Scatter plots : input data (top), FA-realisation (center), PPMT-realisation (bottom)

6 Concluding Comments

The method presented in this contribution appears to adequately address issues raised related to the lack of affine equivariance of standard transformations to multivariate normality. The case study together demonstrates our method's effectiveness in application to real data. It should be noted however that the spread of the transformed distributions depends on the initial scaling factor $_0$, and the number of variables. The range of the transformed variables increases with decreasing scaling factor and for fixed scaling factor, the range for the transformed variables decreases with increasing number of variables. So a tuning of the scaling parameter prior to simulation is essential. Lastly, the current implementation is relatively slow, limiting the size of the data sets that can be handled. Whilst the projection pursuit method does not perform as well as the flow anamorphosis introduced above and adaptation of it by replacing the quantile matching by a different rescaling might also offer a good alternative.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Aitchison, J. (1997). The one-hour course in compositional data analysis or compositional data analysis is simple. In V. Pawlowsky-Glahn, editor *Proceedings of IAMG'97 – The III Annual Conference of the International Association for Mathematical Geology*. pp 3-35.

Barnett, R. M., Manchuk, J. G. and Deutsch, C.V. (2014). Projection pursuit multivariate transform. *Mathematical Geosciences 46* (2), 337–360.

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

Filzmoser, P. and Hron, K. (2008). Outlier detection for compositional data using robust methods. *Mathematical Geosciences 40* , 233–248.

Friedman, J. H. and Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers c-23* (9), 881–890.

Leuangthong, O. and Deutsch, C.V. (2003). Stepwise conditional transformation for simulation of multiple variables. *Mathematical Geology 35* (2), 155–173.

Matheron, G. (1965). *Les variables régionalisées et leur estimation - une application de la théorie des fonctions aléatoires aux sciences de la nature* Masson et Cie, Paris pp 305.

Rivoirard, J. (1984). Une methode d'estimation du recuperable local multivariable. Note 894,CGMM, Mines-Paris Tech pp 10.

Tercan, A.E. (1999). The importance of orthogonalization algorithm in modeling conditionald istributions by orthogonal transformed indicator methods. *Mathematical Geology 31* (2), 155–173.

Tolosana-Delgado, R. (2006). Geostatistics for constrained variables: positive data, compositions and probabilities. Application to environmental hazard monitoring PhD Thesis, Universitat de Girona (Spain) pp 198.

Ward, C. and Mueller, U. (2012). Multivariate estimation using logratios: a worked alternative. In P. Abrahamsen et al, editors *Geostatistics Oslo 2012*, pp 333–343.