Statistical treatment of grain-size curves and empirical distributions: densities as compositions?

R. Tolosana-Delgado[1], **K.G. van den Boogaart**[2], **T. Mikes**[1] **and H. von Eynatten**[1]
[1]Georg-August University, Göttingen, Germany; *raimon.tolosana@geo.uni-goettingen.de*
[2]Ernst-Moritz-Arndt University, Greifswald, Germany.

Abstract

The preceding two editions of CoDaWork included talks on the possible consideration of densities as infinite compositions: Egozcue and Díaz-Barrero (2003) extended the Euclidean structure of the simplex to a Hilbert space structure of the set of densities within a bounded interval, and van den Boogaart (2005) generalized this to the set of densities bounded by an arbitrary reference density. From the many variations of the Hilbert structures available, we work with three cases. For bounded variables, a basis derived from Legendre polynomials is used. For variables with a lower bound, we standardize them with respect to an exponential distribution and express their densities as coordinates in a basis derived from Laguerre polynomials. Finally, for unbounded variables, a normal distribution is used as reference, and coordinates are obtained with respect to a Hermite-polynomials-based basis.

To get the coordinates, several approaches can be considered. A numerical accuracy problem occurs if one estimates the coordinates directly by using discretized scalar products. Thus we propose to use a weighted linear regression approach, where all k-order polynomials are used as predictand variables and weights are proportional to the reference density. Finally, for the case of 2-order Hermite polinomials (normal reference) and 1-order Laguerre polinomials (exponential), one can also derive the coordinates from their relationships to the classical mean and variance.

Apart of these theoretical issues, this contribution focuses on the application of this theory to two main problems in sedimentary geology: the comparison of several grain size distributions, and the comparison among different rocks of the empirical distribution of a property measured on a batch of individual grains from the same rock or sediment, like their composition.

Key words: probability measure, granulometry, size composition, non-parametric representation of probability density

1 Introduction

To measure grain size distributions of sediments, several analytical techniques are available: Laser Particle Sizer (LPS), sieving and settling. In any of these cases, grain size measures are expressed in $\phi = -\log_2(d)$ scale, being d the particle diameter in mm. Typically, in the LPS case one has ~ 100 equally-spaced classes, from $\phi > -1$ up to $\phi \sim 15$. A measurement gives the proportion of particles falling in each of the classes, which is afterwards converted to proportion of volume or mass, by assuming spherical particles of the same density for the whole spectrum. In the interval $(-1, 0)$, measurements are not very reliable: in fact, LPS is good for fine fractions up to $\phi = 13$, where particles become rather colloids and measurements are again quite arbitrary. The grain size curve for coarse grains is measured with other techniques, mainly sieving and settling. For sand, granules and coarser particles ($-1 < \phi < 4$), sieving each ϕ unit is a common choice. For $\phi \in (4, 9)$ the common technique is centrifugation followed by settling.

Numerical treatment of grain size curves has been typically done on the basis of some quantiles read from the empirical cummulative distribution function (ecdf): median, sorting (either as standard deviation, or as range between 5%-95% and/or 16%-84% quantiles), and similar functions to measure skewness and kurtosis (also involving ranges between symmetric quantiles). The goal of such a parametric representation of grain size was originally to allow for a concise yet informative description of the grain size ecdf without having to give a table or a graphic. In this way, comparison between samples was easier, though clearly conditioned to what was being measured by the parameters: these descriptive parameters were based on the normal distribution (mean and sorting) or measured specific departures from it (skewness and kurtosis), and could not describe further issues, e.g. as no parameter measured bimodality, that characteristic was not compared among samples. Alternative treatments have also been developed based on maximum likelihood estimation of some parameters of a previously chosen model, and Kullback-Leibler divergences (Lwin, 2003). *Sediment transport analysis*, put forward by McLaren and Bowles (1985) characterizes the change on the grain-size distribution along a pathway with the quotient of the child against the parent distributions, and study that quotient as characteristic of the process involved, instead of using a parametrization.

The goal of this contribution is to look for a richer set of parameters, beyond the normal distribution and able to be adapted to a specific, desired description. This is done with two aims, namely: (i) to look for a method to reconciliate grain size curves obtained with several methods; and (ii) to compare curves from different individuals, in order to find proximities, groups and evolution patterns between them, in the line of McLaren and Bowles (1985).

Finally, an equivalent situation might arise when measuring any other characteristic of single crystals/grains in a rock/sediment. As a precedent on this line, Sircombe (2000) treated with an Aitchison geometry the age distribution of detrital zircon obtained with U-Pb geochronological techniques: he divided the age span in 49 equally spaced age classes, applied a zero replacement technique to fill empty classes, and then used principal component analysis on the clr transformed data to derive possible sediment transport pathways. The example we will treat here relate to the amounts of Cr/Al and Fe^{2+}/Mg in the spinel crystals of a sediment, instead of the mass of all particles. In this case, the goal will be the derivation of similarities between several stratigraphic units.

2 Method

2.1 The $\mathbb{A}^2$ space of distributions

2.1.1 Preliminary definitions and notation

Let Z be a random variable with support $Dom(Z)$. Let $N(z)$ be a bounded distribution on $Dom(Z)$, with density $n(z) =: n_\lambda(z) = dN(z)/d\lambda(z)$ with respect to the Lebesgue measure $\lambda(\cdot)$, i.e.

$$0 < \int_{Dom(Z)} n(z) d\lambda(z) = \int_{Dom(Z)} dN(z) < +\infty,$$

and giving some probability to each subset of $Dom(Z)$, thus $n(z) > 0, \forall z \in Dom(Z)$. This distribution is considered the *reference distribution*, and will play the role of *origin* of the following space. Take $\mathbb{A}^2(n)$ as the set of possible distributions $F(z)$ on $Dom(Z)$ with square-integrable log-density with respect to $N(z)$,

$$\int_{Dom(Z)} \log^2 \left(\frac{dF(z)}{dN(z)} \right) dN(z) = \mathrm{E}_N \left[\log^2 \left(\frac{dF(z)}{dN(z)} \right) \right] < \infty,$$

where

$$f_N(z) = \frac{dF(z)}{dN(z)} = \frac{dF(z)/d\lambda(z)}{dN(z)/d\lambda(z)} = \frac{f_\lambda(z)}{n_\lambda(z)}$$

is the density of $F(z)$ with respect to $N(z)$. From now on, we will consider $N(z)$—or $n(z)$, being univocally linked—fixed and known, and we will not specify any more that densities are computed with respect to it.

Let f and g be two elements of $\mathbb{A}^2(n)$, and $\mu \in \mathbb{R}$ a real value. We will say that f and g are equivalent in an Aitchison sense, written $f =_A g$, if there exists a unique positive scalar μ fulfilling

$$f(z) = \mu \cdot g(z), \quad \forall z \in Dom(Z). \tag{1}$$

This defines an equivalence class. For those equivalence classes with bounded integral, one can take the representative of the class as the one being a true probability density (i.e., integral one).

2.1.2 Hilbert space structure

The following two operations

$$f \oplus g =_A f(z) \cdot g(z) \quad \text{and} \quad \mu \odot f =_A f^\mu(z), \tag{2}$$

called perturbation and power transformation, define a vector space structure on $\mathbb{A}^2(n)$. The operation

$$\mathrm{clr}(f) = \log f(z) - \int_{Dom(Z)} \log f(z) dN(z) \tag{3}$$

allows the definition of a scalar product on $\mathbb{A}^2(n)$, namely

$$\langle f, g \rangle = \int_{Dom(Z)} \mathrm{clr}(f) \mathrm{clr}(g) dN(z), \tag{4}$$

building a full Hilbert space structure on $\mathbb{A}^2(n)$, according to van den Boogaart (2005).

To interpret these operations, the power operation perfectly describes arbitrary sediment sorting, as it keeps the position of the modes but increases its "peakness", or reduces its dispersion. Regarding perturbation, note that in McLaren and Bowles (1985) approach, the difference between two grain-size distributions would be measured as $g(z)/f(z)$, which is equivalent to (2) without reclosure. Following these authors, if $g(z)$ is the distribution of a daughter sediment of a parent sediment with grain size curve $f(z)$, then

- nett erosion of new sediment would be characterized by a curve $g \ominus f$ with marked negative skewness,
- nett accretion produce a curve $g \ominus f$ with marked positive skewness,
- a dinamic equilibrium between these two generates symmetric $g \ominus f$ curves,
- sediment deposition forms "heavy tailed" $g \ominus f$ curves on those grain sizes which are *not* being deposited,

2.1.3 Exponential families

A K-parametric exponential family is a set of densities such that any of them can be expressed as the product of three positive functions

$$f(z|\varphi_1, \varphi_2, \ldots, \varphi_K) = A(\varphi_1, \varphi_2, \ldots, \varphi_K) \cdot \exp\left[\sum_{k=1}^{K} \varphi_k \cdot T_k(z)\right] \cdot g(z),$$

this is, a function $A(\cdot)$ of the parameters only, a function $g(z)$ only of the random variable, and the exponential of a linear interaction function. Taking $g(z) = n(z)$, one finds that exponential families can be understood as K-dimensional subspaces of $\mathbb{A}^2(n)$, being the vectors of the basis $e_k(z) =_A \exp(T_k(z))$. Then, $A(\cdot)$ can be ignored without risk, as it just closes the density to integral one, i.e. chooses the probability representative of the equivalence class (van den Boogaart, 2005).

2.1.4 Coordinates

Let $\pi_k(z)$ be a series of functions with natural index k, such that $\pi_0(z) \propto 1$ and they are orthonormal with respect to the density $n(z)$, i.e.

$$\int_{Dom(Z)} \pi_i(z) \cdot \pi_j(z) \cdot n(z) dz = \delta_{ij}. \tag{5}$$

These functions allow to define an orthonormal basis of $\mathbb{A}^2(n)$, by taking

$$p_k(z) =_A \exp[\pi_k(z)] \qquad k > 0.$$

The k-th coordinate φ_k with respect to this basis may be computed with (van den Boogaart, 2005).

$$\varphi_k = \langle f, p_k \rangle = \int_{Dom(Z)} \mathrm{clr}(f(z)) \pi_k(z) dN(z). \tag{6}$$

2.2 Three particular cases

2.2.1 The classical densities

Three particular cases will be of interest for our practical applications, all three using ortogonal polynomials satisfying Eq. (5) to induce the basis, and all three generalizing a density from an exponential family. These are described in Table 1, containing:

- the name of the family
- the reference density
- the support of the underlying random variable
- the name of the orthogonal polynomials inducing the basis
- the orthonormal polynomials defining the basis of the subspace in which the conventional family is defined
- the coordinates of the conventional family in this basis
- the dimension of the subspace,
- and the boundary between proper densities and those with infinite integral.

Table 1: characteristics of the three reference models, and their relation to known probability densities.

name	normal	exponential	uniform
$n(z)$	$\frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})$	$\exp(-z)$	$\frac{1}{2}I(z \in Dom(Z))$
$Dom(Z)$	$(-\infty, +\infty)$	$(0, +\infty)$	$(-1, +1)$
polynomials	Hermite	Laguerre	Legendre
$\pi_1(z)$	z	z	—
$\pi_2(z)$	$\frac{1-z^2}{\sqrt{2}}$	—	—
φ_1	$\frac{\mu}{\sigma^2}$	$\lambda - 1$	—
φ_2	$\frac{1-\sigma^2}{\sqrt{2}\sigma^2}$	—	—
dimension	2	1	0
boundary	$\varphi_2 > -\frac{1}{\sqrt{2}}$	$\varphi_1 > -1$	—

For instance, when working with a standard normal distribution as reference and using Hermite polynomials to define the basis, if we restrict our densities to a two-dimensional subspace of $\mathbb{A}^2(n)$ ($K = 2$), then the resulting subspace contains all possible normal densities, as for instance

$$\begin{aligned}
f_N(z|\mu,\sigma) &= \frac{\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\frac{(z-\mu)^2}{2\sigma^2}\right]}{\frac{1}{\sqrt{2\pi}}\exp\left[-\frac{z^2}{2}\right]} = \frac{1}{\sigma}\exp\left[-\frac{(z-\mu)^2}{2\sigma^2} + \frac{z^2}{2}\right] = \\
&=_A \exp\left[-\frac{z^2+\mu^2-2\mu z-\sigma^2 z^2}{2\sigma^2}\right] =_A \exp\left[-\frac{z^2-2\mu z-\sigma^2 z^2}{2\sigma^2}\right] = \\
&=_A \exp\left[\frac{1-\sigma^2}{2\sigma^2} - \frac{1-\sigma^2}{2\sigma^2} - \frac{(1-\sigma^2)z^2-2\mu z}{2\sigma^2}\right] = \\
&=_A \exp\left[\frac{1-\sigma^2}{2\sigma^2} - \frac{(1-\sigma^2)z^2-2\mu z}{2\sigma^2}\right] = \\
&= \exp\left[-\frac{(1-\sigma^2)(1-z^2)+2\mu z}{2\sigma^2}\right] = \exp\left[\frac{1-\sigma^2}{\sqrt{2}\sigma^2}\frac{1-z^2}{\sqrt{2}} + \frac{\mu}{\sigma^2}z\right] = \\
&=_A \frac{\mu}{\sigma^2}\odot e^{\pi_1(z)} \oplus \frac{1-\sigma^2}{\sqrt{2}\sigma^2}\odot e^{\pi_2(z)} = \varphi_1 \odot e^{\pi_1(z)} \oplus \varphi_2 \odot e^{\pi_2(z)}.
\end{aligned}$$

(The notation $=_A$ is used when the two members of the equality are equivalent in Aitchison sense, that is: from one to the other an element was added or removed, which is a multiplicative function of the paramerers only, not of the variable z). Note that, in this case we know that the result will only be a probability density (i.e., with finite integral) iff $\sigma^2 > 0$, which implies that $\varphi_2 > -1/\sqrt{2}$. On the contrary, $\varphi_1 = \mu/\sigma^2$ can take any value. Thus, the set of densities is a 2-dimensional half-subspace of $\mathbb{A}^2(n)$, with the other half containing measures with infinite integral.

2.2.2 Densities of higher dimension

If we allow more polynomials in the basis, resulting densities will no longer be of the normal family, and even these relationships between means, variance and coordinates will not be satisfied any more. In this case, results will only be probability densities if the maximum degree of the polynomials (K, dimension of the subspace) is even, and its coordinate φ_K is negative: these conditions ensure that the log-density decreases towards both $\pm\infty$. Similar conditions must be imposed in the exponential family to ensure proper densities: as the density is defined for $z > 0$, one has to ensure that the resulting polynomials will decrease towards $+\infty$, that is its highest degree term must have a negative coefficient, be it even or odd. Table 2 summarizes these conditions.

Interestingly, the range of shapes that can be displayed by the resulting densities is much broader than those of their parent classical distributions (Fig. 1),

Table 2: boundaries on the parameter space between proper and improper densities, for the three reference models.

name	normal	exponential	uniform
maximum K	even	any	any
φ_K	< 0	< 0	no condition

- the *maximum* number of (derivable) modes is equal to $K/2$ (or the biggest integer below it), though with the exponential one *can* have another maximum extreme at $z = 0$, and with the uniform the same can happen in any one or both extremes ($z = \pm 1$),
- in the normal and uniform case, the skewness is fundamentally controlled by the odd coordinates, as even polynomials contain only even degrees and have thus always symmetric contributions,
- polymodality in the normal and uniform cases is controlled by the difference between the coefficients of the highest degree and the following even coefficient.

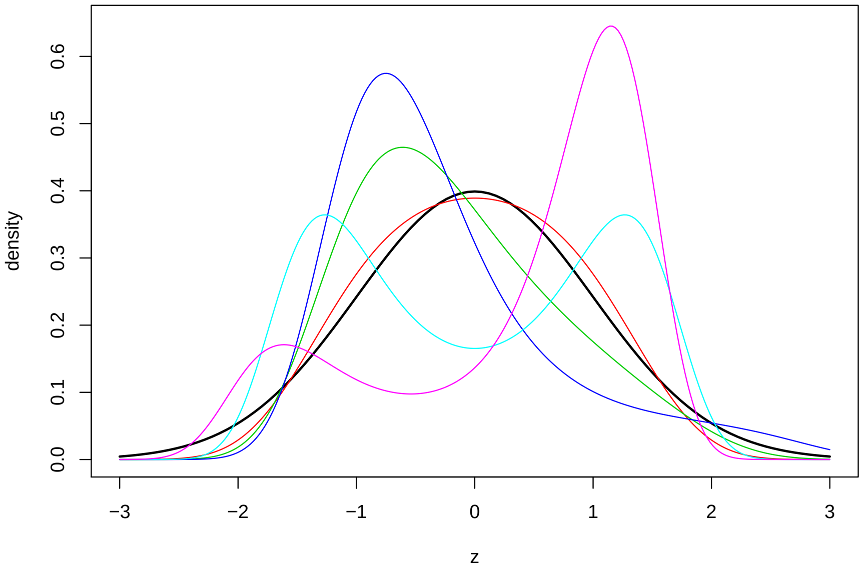

Figure 1: Several examples of densities of the exponential family formed by the normal density (bold black line) and the first 4 Hermite polynomials, with the following coordinates:

colour	φ_1	φ_2	φ_3	φ_4
red	0.00	-0.50	0.00	-0.50
green3	0.00	-0.50	0.50	-0.50
blue	0.00	-0.50	1.00	-0.50
cyan	0.00	-0.50	0.00	-1.50
magenta	0.00	-0.50	-1.00	-1.50

2.3 Discrete approximations to coordinate computation

With empirical data, one will almost always have just a discretized version of the densities or distributions, not the functions themselves. These can be for instance,

- a kernel-density estimation of the distribution of a sample
- the laser particle sizer measurement of the granulometric curve of a sediment
- a kernel-density approximation of the sieving measurement of the granulometric curve of a sediment

Equation (6) will thus be very rarely useful to obtain the coordinates of these functions. A first idea is to try a **quadrature approximation**. Let $z_1, \ldots, z_M$ be a set of M equally spaced points in $Dom(Z)$, and $f_1, \ldots, f_M$ and $n_1, \ldots, n_M$ the values of the target density $f_m \propto f(z_m)$ and the reference density $n_m \propto n(z_m)$ on these points, the last one forced to sum up to one. Then we should have

$$\varphi_k \approx \sum_{m=1}^{M} \operatorname{clr}(f_m) \cdot \pi_k(z_m) \cdot n_m,$$

with

$$\operatorname{clr}(f_m) = \log \frac{f_m}{n_m} - \sum_{i=1}^{M} n_i \log \frac{f_i}{n_i} \tag{7}$$

But we should also verify that the orthonormality conditions hold (Eq. 5), i.e.

$$\sum_{m=1}^{M} \pi_j(z_m) \cdot \pi_k(z_m) \cdot n_m \approx \delta_{jk}.$$

The error of this approximation is quite noticeable. For instance, the first 6 Lagrange normalized polynomials, orthogonal with respect to the uniform density in the interval $(-1, +1)$, using 512 discretization points give the following approximate scalar products

	π_1	π_2	π_3	π_4	π_5	π_6
π_1	0.5020	-0.0000	0.0045	-0.0000	0.0057	-0.0000
π_2	-0.0000	0.5039	-0.0000	0.0066	-0.0000	0.0080
π_3	0.0045	-0.0000	0.5059	-0.0000	0.0087	-0.0000
π_4	-0.0000	0.0066	-0.0000	0.5079	-0.0000	0.0108
π_5	0.0057	-0.0000	0.0087	-0.0000	0.5100	-0.0000
π_6	-0.0000	0.0080	-0.0000	0.0108	-0.0000	0.5120

Non-bounded reference distributions produce higher errors, and they become higher with the degree of the involved polynomials. Possible reasons of that behaviour are that more coefficients to approximate imply more error, and also that higher degree polynomials have more fluctuations (and are thus worse approximated by piece-wise linear splines as is implicitly done by the quadrature). Since this matrix of scalar products is not exactly the identity, the functions used are not exactly orthonormal with respect to the discretized density $\{n_m\}$. Thus a proper computation of the coordinates should have this obliqueness into account: computation of the coefficients should be done by inverting this scalar product matrix. Notice that, for a vector of coordinates $\boldsymbol{\varphi} = [\varphi_1, \ldots, \varphi_K]$ we could recover its associated density with

$$f(z) = \bigoplus_{i}^{K} \varphi_i \odot p_i(z), \tag{8}$$

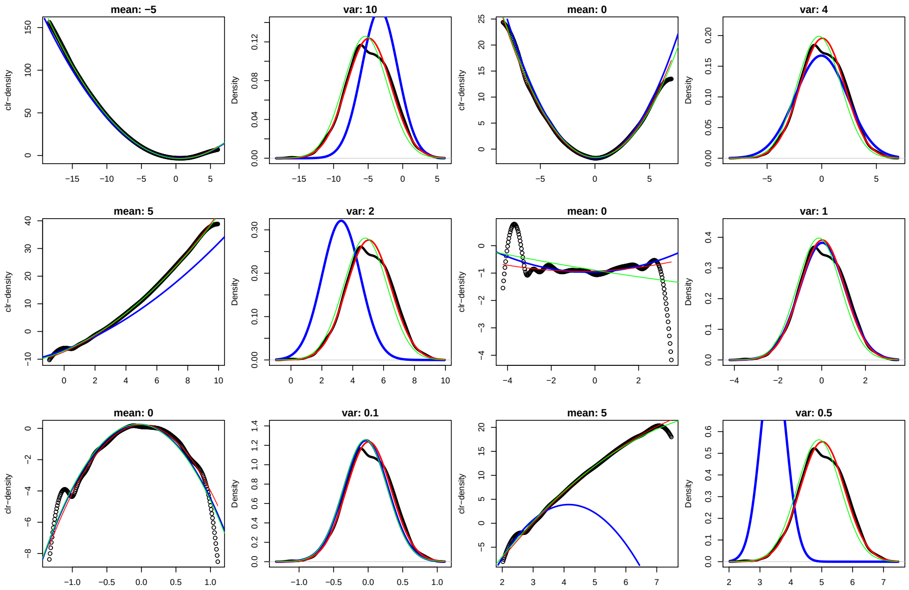

Figure 2: Estimation of coordinates for a normal distribution, with weighted regression (blue), with conventional regression (green), and with the relationships with the moments (red), applied to a kernel density estimate (black) obtained from the same random sample of size 1000 (for each case, conveniently scaled and translated according to the variance and mean given in the pairs of plots).

with the vector $p_i(z)$ the exponential of the appropiate i-degree polynomial defining the basis. By multiplying it with another vector of the basis, one gets

$$\langle f(z), p_j(z) \rangle = \sum_i^K \varphi_i \cdot \langle p_i(z), p_j(z) \rangle,$$

Take the vectors of the values of the involved functions on the M discretization nodes, conveniently clr-transformed through the discete version of (Eq. 3), and denote them by $\mathbf{p}_i = [\pi_i(z_1), \ldots, \pi_i(z_M)]$ and $\mathbf{f} = [f^*(z_1), \ldots, f^*(z_M)]$, where $f^*(z_m) = \text{clr}(f(z_m))$. Let us arrange all $\mathbf{p}_i$ in a $M \times K$ matrix $\mathbf{P}$. Finally, define weights $w_m \propto n(z_m)$ and summing up to one, and arrange them in a diagonal matrix $\mathbf{W}$. The previous expression with scalar products can be expressed for all $p_j(z)$ simultaneously as

$$\mathbf{P}^t \cdot \mathbf{W} \cdot \mathbf{f} = \mathbf{P}^t \cdot \mathbf{W} \cdot \mathbf{P} \cdot \boldsymbol{\varphi}, \tag{9}$$

which coincides with the equation of **weigthed linear regression**, identifying the explanatory variables as the polynomials $\mathbf{P}$, and the explained variable as the studied function $\mathbf{f}$. In other words, the slight departures from orthonormality shown can be compensated by using weighted linear regression to estimate the coordinates instead of directly projecting the function onto the basis in use.

Finally, there is a third way to compute the coordinates, only valid in case of restricting our study to the classical densities, that is, 2 coordinates for a normal, 1 coordinate for an exponential. In this case, the relations between the **conventional parameters** (mean, variance, scaling) and the coordinates can be found in Table 1.

Figure 2 compares the results of these two last methods (weighted regression and moment estimation), for several combinations of means and variances. The method of plugging the empirical moments in the equations for the coordinates provides excellent results (and quickly), but it is not as interesting (there are other standard ways of comparing normal distributions) and is not valid for other general cases. On the contrary, weighted regression is extremely sensitive to the location parameter of the distribution. Using weights is nevertheless interesting because they focus the fit on an area around the mean, and downweight the tails, prone to stronger fluctuations. However, the farther the true mean from the origin is, the worse the regression becomes, because the "interesting" part of the studied distribution falls in the tail of the reference distribution, the standard normal density. Here the concept "farther" should be evaluated in a Mahalanobis sense, with respect to the variance of the studied distribution and the reference. In conclusion, it seems interesting to allow the reference distribution to be adapted to each particular application. Nevertheless, using weighted regression should be done with caution: our experiments (not reported here) led us to conclude that the system becomes easily singular, specially when including high order polynomials.

2.4 Proposed algorithm and R functions

With the preceding theoretical and practical considerations, the following algorithm is proposed to statistically treat densities. We implemented some of its steps in several functions in `R`, following the conventions and classes of the package "`compositions`" (van den Boogaart and Tolosana-Delgado, 2008). Given that "`density`" is an existing object class in `R`, we generated an *Aitchison*-geometry-density class under the name of `adensity`. We created some methods for treating them in the generic functions already existing, like `cdt` and `idt` (wrappers around clr and ilr transformations), but also some functions specially defined for density treatment. These functions are mentioned in the following points.

(1). Look at the several densities to be compared, and choose the support where the density will be studied (minimum and maximum values for z, number of discretization points). To obtain a density from a data set, use `adensity(z)`, a wrapper on the standard function `adensity(z)` for kernel estimation. To obtain a density from sieving data, use `sieve2adensity` (with arguments `x=` sieve ϕ values and `y=` mass on each sieve).

(2). Choose a reference distribution, the highest degree of the polynomials, and a scaling of the support (function `gsi.ilrBase.adensity` does it explicitly, but `idt.adensity` calls this one implicitly):

(2.1) the reference density can be a normal, an exponential and an uniform one

(2.2) the highest degree K must be even for the normal case, and can be freely chosen for the other two

(2.3) the scaling

$$z^* = \frac{z-a}{b}$$

of the support is aimed at adequately focusing the reference distribution, and it can be done automatically to ensure that the reference density has only a probability α=`alpha` outside the support:

	a	b
normal	$\frac{\min(z)+\max(z)}{2}$	$\frac{\max(z)-\min(z)}{2q_{\alpha/2}}$
exponential	$\min(z)$	$\frac{\max(z)-\min(z)}{q_{\alpha}}$
uniform	$\frac{\min(z)+\max(z)}{2}$	$\frac{\max(z)-\min(z)}{2}$.

Here q_α represents the α quantile of a standard normal or an exponential distribution (with $\lambda = 1$).

(3). Compute the coordinates of each density with respect to the chosen basis (`idt.adensity`), with the following sub-steps:

(3.1) scale the support as mentioned before (and use z^* from now on),

(3.2) remove those nodes where the discretized density is zero, (3.3) obtain the weights w_i as proportional to the reference density evaluated at the remaining nodes, and summing up to one

(3.4) apply the clr transformation for a discretized density (Eq. 7) using the weights w_i instead of n_i,

(3.5) use weighted regression (Eq. 9) to explain the clr-transformed density as a linear combination of the first $K + 1$ orthonormal polynomials with respect to the chosen density (non-weighted regression is used instead if the optional argument `weighted=FALSE` is given to `idt.adensity`),

(3.6) remove the coefficient of the constant; the remaining coefficients are the estimates of the coordinates (an object of class `adensity.rmult`).

(4). Apply the desired statistical method to the coordinates.

(5). For those results describing a density, apply the coordinates to the basis to recover that density (`idt.inv.adensity.rmult` to obtain a `adensity` object, or `lines.adensity.rmult` to plot it). This is done as follows:

(5.1) at all M scaled nodes, evaluate the orthonormal polynomials (not including the constant $\pi_0(z^*) \propto 1$)

(5.2) multiply the polynomials by the coefficients

(5.3) add the logarithm of the reference density evaluated at the nodes

(5.4) take exponentials,

(5.5) rescale the support, to recover the original $z = a + b \cdot z^*$

(5.6) close the resulting discretized density to sum up to one, conveniently weighted by the node interdistance $(\max(z) - \min(z))/(M - 1)$, as would correspond with a histogram.

3 Applications

3.1 Sand size distribution of the Darss Sill

This example deals with a typical reference data set of grain size problems: the Darss sill data set. The Darss Sill is situated between the Danish isles of Falster and Møn in the northwest and the Darss peninsula in Germany in the southeast, at the entrance of the Baltic Sea. Lemke (1997) reports the main geologic and physiographic conditions controlling the distribution of sediment on this structure. The data set contains 1281 sandy surface samples, which were dried and sieved into eight weight-percent size fractions, from less than 63 μm (silt and finer) to over 2000 μm (gravel and coarser).

This data set was the focus of a session in IAMG'97 (Pawlowsky-Glahn, 1997), where reseachers presented several approaches. Among them, two are linked to the approach of this contribution, though in different ways. Martín-Fernández et al. (1997) took the eight grain size categories as parts of a composition, amalgamated four of them (from gravel to medium sand) and applied the additive log-ratio transformation to the resulting 5-part composition, after a zero substitution of the remaining zero components. These log-ratio transformed values were the data to a parametric classification scheme, generating 7 groups. On the other hand, Tauber (1997) suggested to treat them as a continuous distribution, and fitted an analytic approximation of the normal cdf to the empirical cdf of the grain size. This analytic function depended on a location parameter (*median*) and a scale parameter (*sorting*). These two parameters were then used as "data" in the subsequent analysis (clustering, regionalization, etc.), concluding that the grain size distributions showed more a continuous pattern than groups.

Both approaches have interesting advantages and some inconvenients. The log-ratio approach requires amalgamation and zero substitution, which respectively removes some important sedimentologic information and generates spurious clusters (Tauber, 1999). But the median-sorting

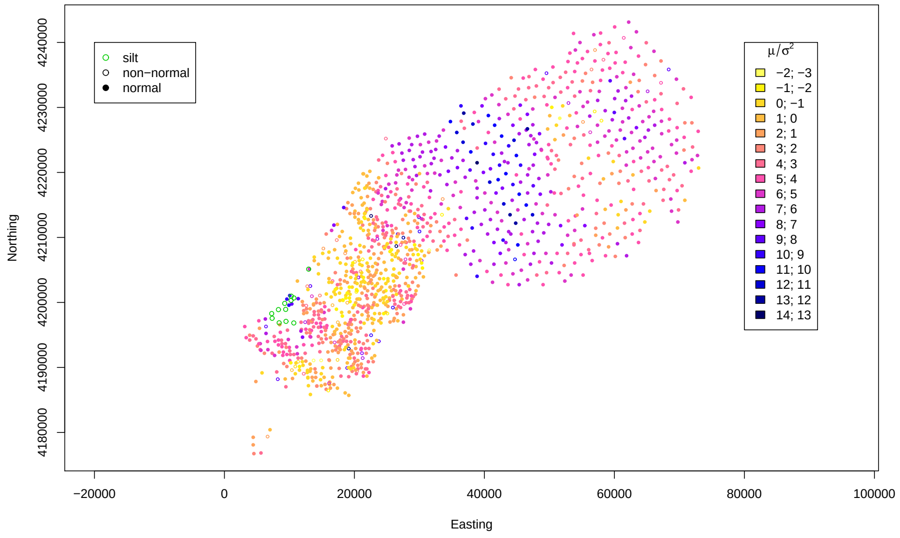

Figure 3: Map of the first coordinate of the grain size distribution of the Darss sill samples (Baltic Sea). Silt samples are singularized with green circles (correspond to groups 8 and 9 of non-normal samples). Most of the samples admitted a representation with k=2 (labelled as "normal"), whereas some few present important skewness/kurtosis (labelled as "non-normal"; for them, the expression of the first coordinate as a function of the moments is *not* valid).

approach is limited to normal distributions, and would yield very different results if we change the parametrization of the curve, which in the end is something arbitrary; e.g. we could work with standard deviation instead of sorting (they are proportional for normal curves), with variance, with their inverses or with their logarithms. Any of these parametrizations is valid and meaningful in a context, but they would not lead to the same results. In contrast, Martín-Fernández et al. (1997) is "non-parametric", and can incorporate non-normal distributions. Nevertheless, both approaches are "spurious-correlation-free".

The approach proposed in this contribution builds a bridge beween them: we found that by generalizing the log-ratio approach to continuous densities, the two first coordinates to describe a normal distribution provide a "natural" parametrization. The euclidean distance between these coordinates is (approximately) isometric to the log-ratio distance compatible with the Hilbert space structure [Eqs. (2-4)]. And if one has non-normal distributions (skew or kurtotic), adding more coordinates one can describe them in the same consistent way. Moreover, we do not need to replace zeroes of the composition, as they are removed in step (3)3.2 from the computation.

To treat the data (previously recasted to a density by `sieve2adensity` function), we used a normal distribution as reference, scaled between ϕ -4 and 7 (probability 1% outside this interval), and $k = 4$ Hermite polynomials as basis. Most of the curves, nevertheless, were not satisfactorily described with four coordinates, as the last one got a positive coefficient (i.e., they would not be bounded densities). Therefore, for them we switched to $k = 2$, thus a normal distribution actually: these are mapped in Figure 3, by using the first coordinate, which is a sort of coefficient of variation (Table 1).

For those "non-normal" data, a Ward cluster analysis was used to try to find similar patterns of deviation from normality. The dendrogram of Figure 4 shows the dendrogram, and Figure 5 the spatial distribution of each cluster. This last figure also shows the several grain size curves of each group and their average. This average is the grain size curve having as parameters the average of the parameters of the curves within the group.

The first group suggests bimodality, a mixture of fine sand and clay/silt (with missing very find

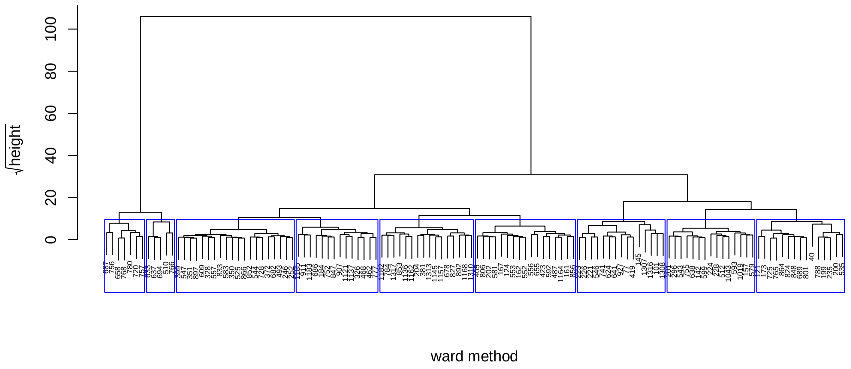

Figure 4: Dendrogram of Ward cluster analysis of the non-normal grain size curves. Note that the vertical scale is represented as its square root.

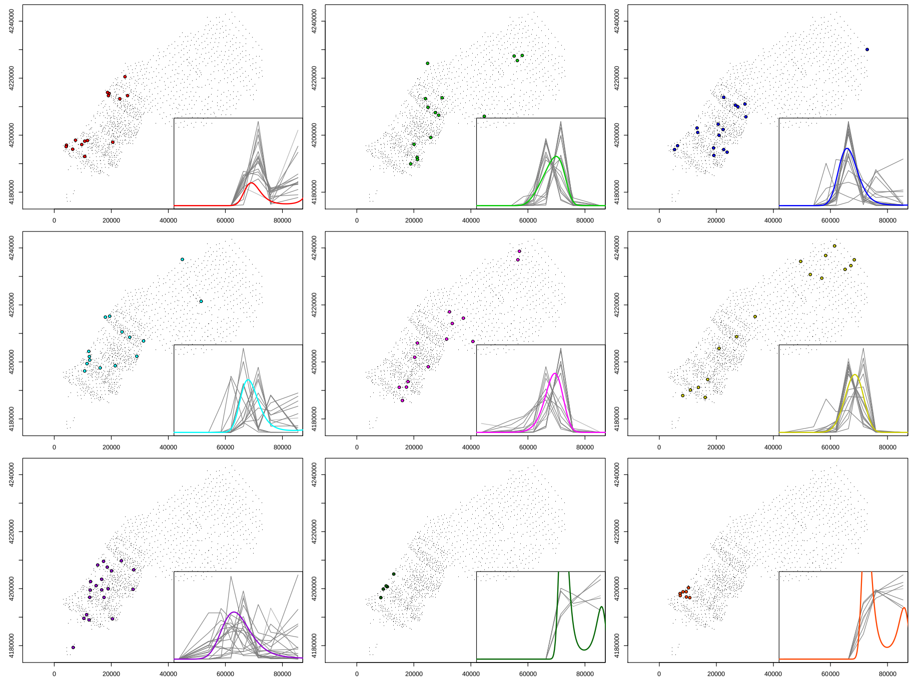

Figure 5: The 9 groups of non-normal grain size curves. For each group, the spatial distribution is shown (main plot), as well as the ϕ distribution of all the data, together with the average distribution (that one with coordinates the average within the group).

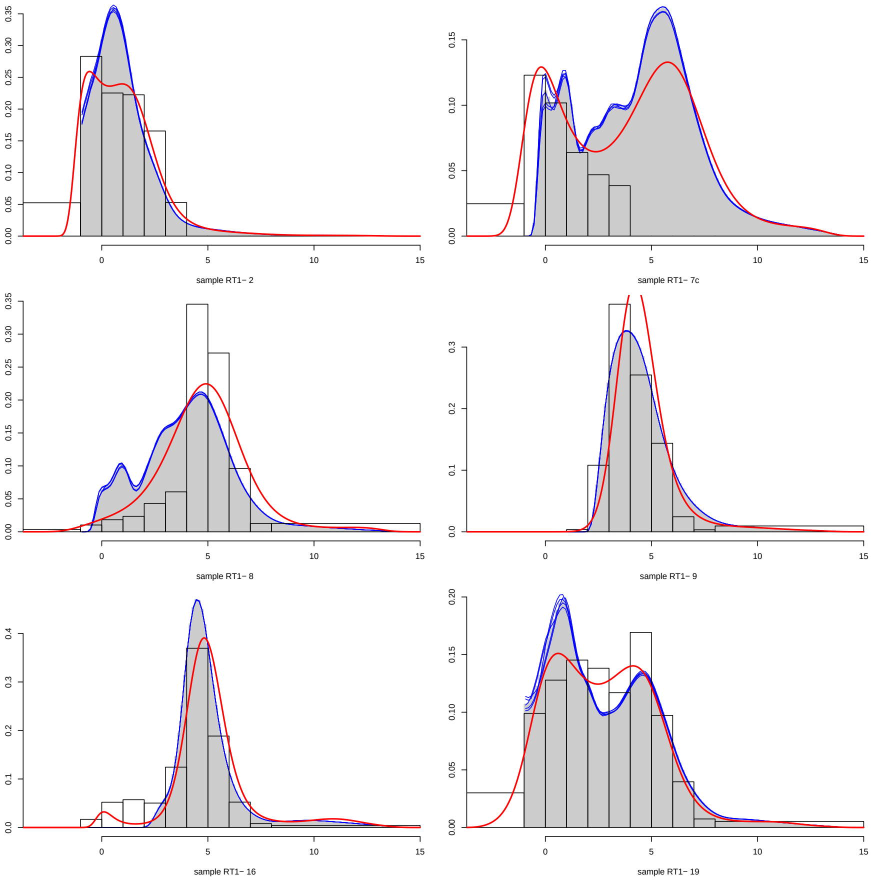

Figure 6: Grain size measurements of four sediment samples, with two techniques: laser particle sizer (blue lines for each of the 5 measurements, gray area for their average), and sieving plus centrifugation-settling (histogram). The red line represents the estimated reconciled curve.

sand). The second and third groups are fine sands, with asymmetry to the right and the left respectively. The fourth group is very similar to the third, but with a sharper lack of medium sand. The fifth and the second are also very similar, though the second is clean of coarser sand categories which occur in the fifth. The sixth and seventh are unclear groups, with very different distributions respectively showing kurtotic and skew averages. Finally the eigth and ninth groups are silt-rich samples, where the method (focusing on the sand domain as it does) failed to correctly fit a curve, and identified a very fine sand mode and a silt mode: these samples are better left behind.

3.2 Reconciling measurements from different methods

In this application, the problem is to take two sets of measurements of the grain size curve, with different support, different precision and different optimal domains, and try to find a global grain size curve accounting for all this information. In the present case (Fig. 6), we have LPS measurements in the range $\phi \in [-1, 14.61]$ split in 116 equally-spaced classes, and another set of combined sieving/settling measures covering the range $\phi \in (-2, 8)$ in units plus a last residual class for grain sizes in $[8, \infty)$. Let us denote these two sets of supports (classes) as $\{x_i, i = 1, \ldots, 116\}$ and $\{y_i, i = 1, \ldots, 11\}$, and the proportion of grains falling in each class as $\{f_i = f(x_i), i = 1, \ldots, 116\}$ and $\{g_i = g(y_i), i = 1, \ldots, 11\}$.

The idea is to find a curve like (8), which lies at a minimum average distance from both LPS and

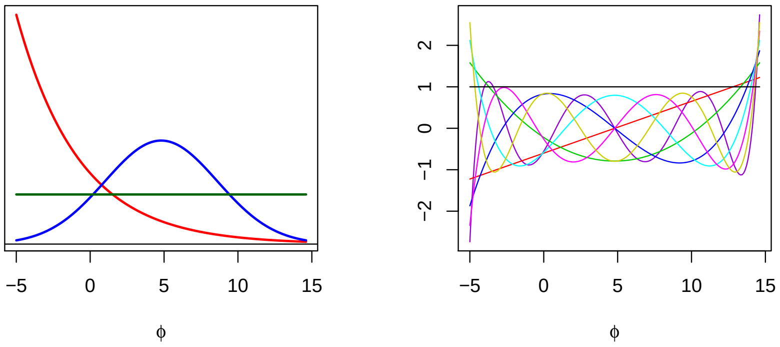

Figure 7: (left) Comparison of the three possible reference curves, scaled to contain at least 99% of probability in the sampled domain. (right) Scaled Legendre orthonormal polynomials used in this example.

sieving/settling measurements. The first step is to choose a reference distribution, giving weights to each x_i or y_i, thus it can incorporate our knowledge on which methods are optimal for which subdomains of the ϕ spectrum. For instance, we know that LPS data from the right part of ϕ should not be trusted: if we take an exponential reference, the result would be the opposite, as those are just the points where the exponential density is higher (see Fig. 7). Both the normal distribution and the uniform could be reasonable choices. In this case, we take the uniform, to be fairer with the sieving/settling data against the LPS.

Then we use $k = 10$ Legendre polynomials $L_k(x)$, orthonormal with respect to the uniform distribution. By evaluating each of them at the sampling nodes, x_i and y_i, we obtain $k+1$ explanatory variables (including the constant $L_0(x) \equiv 1$), used in a weighted linear regression procedure to explain the clr-transformed values of $\{f_i\}$ and $\{g_i\}$. Regarding the constant, remember that its coefficient (the intercept) will be ignored, as it just ensures a sum of zero for this particular clr values: in other words, each of the two data sets should be allowed to have different intercepts but the same coefficients. This is readily obtained by adding another explanatory variable equal to zero for x_i and to one for y_i. And if we had more methods to reconcile, we would add one of these auxiliary intercepts for each of them. The final system is thus:

$$\underbrace{\begin{pmatrix} 0 & 1 & L_1(x_1^*) & \cdots & L_k(x_1^*) \\ 0 & 1 & L_1(x_2^*) & \cdots & L_k(x_2^*) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & L_1(x_{116}^*) & \cdots & L_k(x_{116}^*) \\ 1 & 1 & L_1(y_1^*) & \cdots & L_k(y_1^*) \\ 1 & 1 & L_1(y_2^*) & \cdots & L_k(y_2^*) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & L_1(y_{11}^*) & \cdots & L_k(y_{11}^*) \end{pmatrix}}_{\mathbf{P}} \cdot \underbrace{\begin{pmatrix} \varphi_0' \\ \varphi_0 \\ \varphi_1 \\ \vdots \\ \varphi_k \end{pmatrix}}_{\boldsymbol{\varphi}} = \ln \underbrace{\begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_{116} \\ g_1 \\ g_2 \\ \vdots \\ g_{11} \end{pmatrix}}_{\mathbf{f}},$$

with the following practical comments:

- the support points are rescaled following step (2)2.3 of the general procedure (page 9);
- each points from a LPS measurement has the same weight; to give the sieving/settling data a similar influence while capturing the asymmetry in the reliability of the sieves, we give the first sieve a weight of 2.5×11, the second a weight of 2.5×10, etc., and the eleventh category a weight of 2.5×1; this amounts to giving LPS a total weight of 116 and to sieving/settling 2.5×65

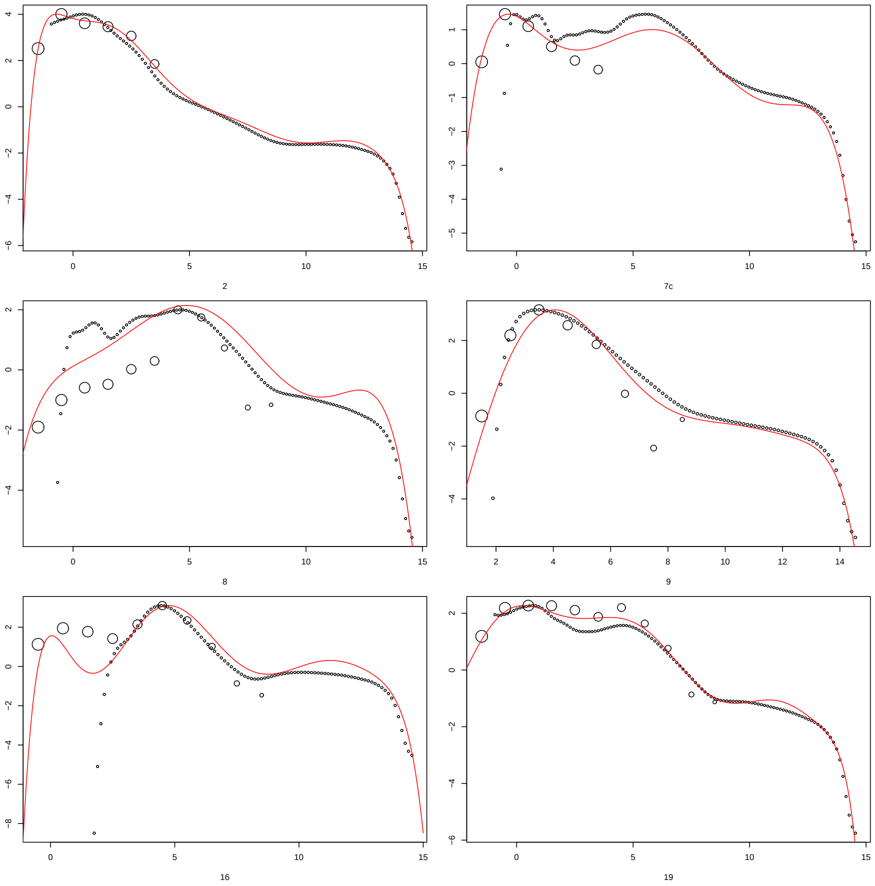

Figure 8: Grain size measurements of four sediment samples, expressed in clr scale, but normalized to the same upper value (instead of to zero sum), in order to enhance the comparison of the three data sets. Big circles correspond to data coming from sieving/settling, and small circles to data from LPS. Their areas are proportional to the weights given to each point. The red line represents the estimated reconciled curve, obtained with Legendre polynomials of highest degree (from left to right, top to bottom): 10, 6, 6, 8, 8, 10.

- if some of the f_i or g_i are zero, the corresponding row of both $\mathbf{P}$ and $\mathbf{f}$ matrices is erased and the weights reclosed;
- if one is using a reference density $n(x)$ different from the uniform one, each f_i and g_i must be replaced by their relative densities, namely $f_i/n(x_i^*)$ and $g_i/n(x_i^*)$.

The result is shown as a red line in both Figures 6 and 8, respectively showing the original f_i and g_i and their clr-transformed values, together with the resulting fitted model.

3.3 Cr-spinel distribution "facies"

Spinel is a part of a quite continuous series of igneous isometric/cubic minerals with formula XY_2O_a where X is bivalent (Fe^{2+}, Mg or Mn) whereas Y is typically trivalent (Al^{3+}, Cr^{3+}, Ti^{4+} and Fe^{3+}). Actual spinel composition depends on its host rock, a fact that gave rise to its widespread use as a petrogenetic indicator mineral (see Barnes and Roeder, 2001, for a review). In mantle rocks, such as peridotites, spinel Ti<0.2 wt% (Kamenetsky et al., 2001), and the Cr/Al ratio is an approximative measure of the degree of partial melting the rocks underwent. Cr/Al ratios are thus low in lherzolites and high in the more depleted harzburgites (Dick and Bullen 1984). Spinel crystallized from magmatic melts has typically TiO_2 >0.2 wt%. Such rocks constitute the majority of ophiolite suites, which play an important role in establishing tectonic models for mountain building; both directly, and indirectly through the study of detrital spinels in sediments derived from the ophiolites. Interest will be in this case to compare the empirical distributions of the proportions of the major end member molecules $FeCr_2O_4$ and $MgAl_2O_4$ of detrital Cr-spinel. The samples represent several Alpine basins in the Dinaride and Carpathian mountain belts. The goal of the approach is to reveal similarities between samples, which might be deemed as "probability-facies" and possibly be used to derive the relative contributions of each of the three rock types to the sediment.

Traditional analysis of Cr-spinel grains is done on the space of the so-called *Mg*-index vs. *Cr*-index, these are the proportions of Mg in $\{Mg, Fe^{2+}\}$ subcomposition and the proportion of Cr in $\{Cr, Al\}$ (Fig. 9). In this space, most of the data fall on the diagonal between the end-member molecules $FeCr_2O_4$ and $MgAl_2O_4$, due to the fact that a significant amount of chemical variation in most Cr-spinels can be readily described in terms of these two molecules alone. At this stage, low-Ti (mantle) and high-Ti (magmatic) spinels need to be treated separately, due to their different origin (Barnes and Roeder, 2001; Kamenetsky et al., 2001).

Due to the petrogenetic significance of the Mg+Cr ↔ Fe^{2+}+Al cation substitution, our data are projected onto this exchange line, and we work with the variable:

$$x = \frac{\frac{Cr}{Cr+Al} + \frac{Fe}{Fe+Mg}}{2}.$$

This is applied a logit transformation, and a kernel density estimation is used for all grains coming from the same tectono-stratigraphic unit, after removing the outliers from each of them. In this way, 41 densities are obtained, and then each one is expressed in $k = 10$ coordinates with respect to a normal distribution and Hermite polynomials. This procedure is applied only to the low-Ti spinels (mantle). At this point, 3 samples were removed due to their extreme character ("Bosnian Flysch / Ugar", "X-Kruja / ALB" and "XN-Krk ML"). To account for the high-Ti (magmatic) spinels in the remaining samples, we derive the proportion of cristals with high Ti from each unit, compute its logit transformation and rescale the result, in order to give it the same standard deviation as the average of the 10 coordinates of the low-Ti part. This provides a data set with 11 variables and 38 cases.

A Ward cluster analysis (Fig. 10) revealed two main groups, two small groups and 2 atypical samples (apart from those 3 removed from the beginning)

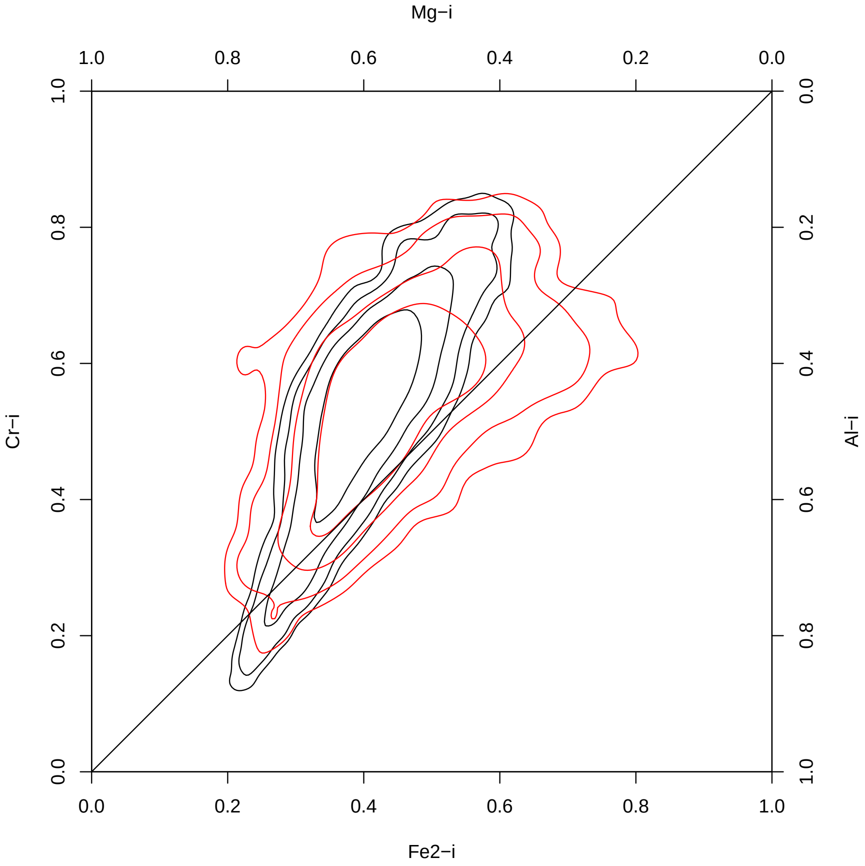

Figure 9: Distribution of Cr-spinel single grain compositions, for high-Ti (red, $> 0.2\%$) and low-Ti (black, $\leq 0.2\%$) crystals. Contours are isodensity curves enclosing 50, 75, 90 and 95% of the data. The density was obtained using a 2D-kernel estimation in the logit-logit transformed space.

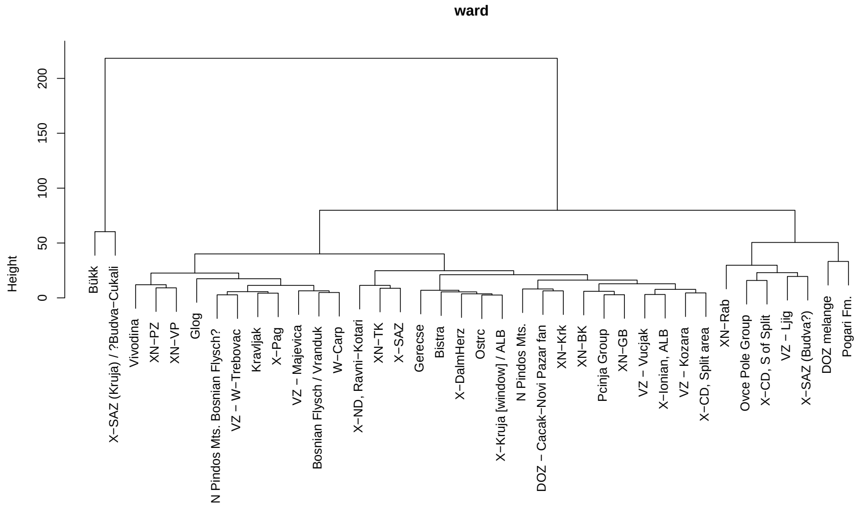

Figure 10: Cluster analysis of the stratigraphic units, using the coordinates of the distribution of $FeCr_2O_6$ vs. $MgAl_2O_6$ proportion, as well as the proportion of low-Ti to high-Ti grains (in log scale). Three samples (“Bosnian Flysch / Ugar”, “X-Kruja / ALB” and “XN-Krk ML”) were completely removed, due to its extreme outlying character.

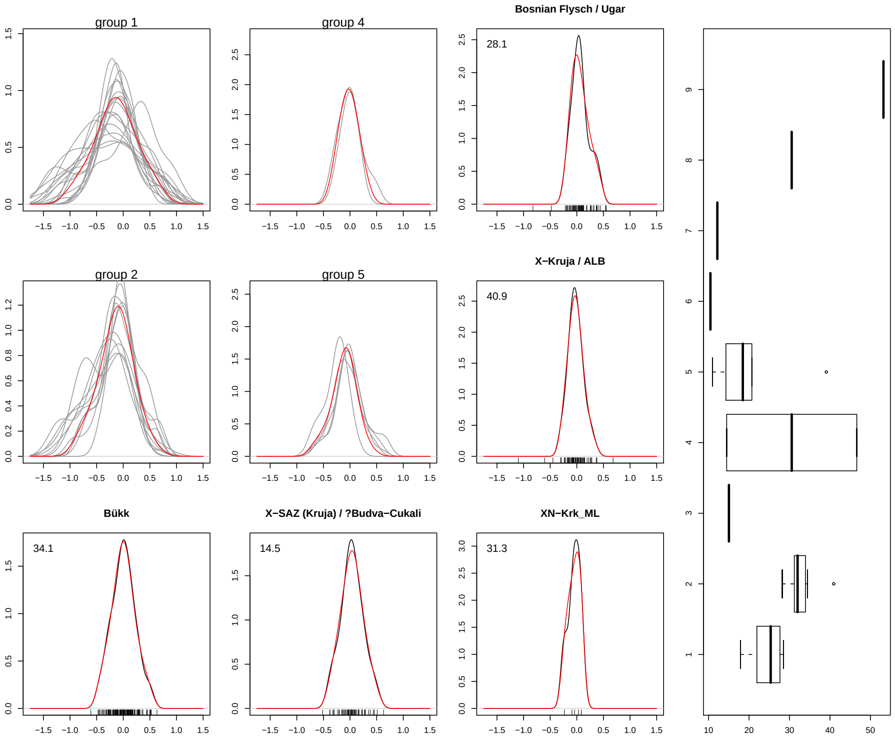

Figure 11: Characterization of the several groups (and single atypical samples) obtained with cluster analysis of the stratigraphic units. For each group, the kernel density estimates of all the samples is represented in grey, and their average (in the coordinate scale) is portrayed in red. For each atypical sample, we represent the kernel density and its coordinate simplification, together with the data set (as a rug in the x axis) and the proportion of high-Ti crystals in the sediment (number at the upper left corner, in %). Finally a box-plot of these high-Ti crystal proportions for each group is also included.

In basins with ophiolitic detritus, provenance of Cr-spinel grains has been traditionally assessed mainly using the range of Cr in the (Cr, Al)-subcomposition, the TiO_2 content, and the calculated Fe^{3+}/Fe^{2+} ratio, by visual estimation from discrimination diagrams. However, Alpine ophiolite complexes are petrologically often extremely inhomogeneous on a map scale (Dick and Bullen, 1984), which is also reflected by the composition of spinels they would release into the sediment. Obduction and subsequent tectonics may further rearrange individual spinel-bearing magmatic and mantle-derived members of the ophiolite suite. Upon tapping by the drainage system, sediment composition will thus largely depend on this previous history of the ophiolite suite. Distribution of detrital Cr-spinel chemical parameters therefore reflects the net effect of oceanic litospheric evolution, obduction, subsequent tectonics and the extent of the drainage, rather than the "ophiolite type" alone. We propose that comparison of sediments should rely on the integrative examination of the useful parameters as presented here. We suggest that any similarity between samples is not necessarily an indication of "palaeogeographic connection" but simply a result of a similar sedimentary response whenever the above evolutionary steps share a common history.

The significance of our approach is twofold. Results reveal that, first, several stratigraphic units, previously tentatively correlated based on geographic proximity, age, lithofacies and tectonic considerations, indeed received detritus of quite similar "final" composition. For example, similarity between the units *DOZ mélange* and *Pogari Fm.*, as well as between the *Gerecse*, *Ostrc* and *Bistra* formations, is corroborated by field and stratigraphic relations. Also, the observed dissimilarity between the *DOZ mélange*e and *Vranduk* are supported by independent geological data.

Second, there are notable differences between some stratigraphic units in the cluster dendrogram (Fig. 10), though they were previously considered as "related" based on some of the above criteria. An evaluation of this technique usefulness to sediment provenance analysis still presents some "geologic" caveats, being the role of the Ti-threshold an important one. For example, let us assume that the units "*X-CD, S of Split*" and "*X-CD, Split area*" are related, based on the similar compositional range of their low-Ti spinels, and the position of local maxima of the polynomial functions, both being around -1 and -0.2 (in logit scale). It is possible that a choice of the TiO_2 threshold different from the actual one would result in a different data distribution. In other words, coordinates of the variation of Mg in the (Fe^{2+}, Mg)-subcomposition, being similar to the Fe-Ti trend of Barnes and Roeder (2001), should also be included in the cluster analysis in future.

The distribution of Ti-rich and Ti-poor spinels in the units, shown in conjunction with the cluster analysis results illustrates the importance of the distribution alone on the eventual separation of the units (Fig.11). For example, groups 1 and 2, the members of which having small Ward distances, differ in the ranges of their logit Ti-functions. A possible interpretation of this pattern is that units in group 2 were supplied by detritus from an ophiolitic source where a higher proportion of magmatic units were subject to erosion, while the composition of the mantle-derived members was comparable for both units. It is recalled that obducted ophiolites are typically dissected along petrological boundaries, thus pursuing this approach may add considerable precision to source area assignment and tectonic reconstructions in future.

4 Closing remarks

This work is a preliminary yet promising application to the theory of "continuous compositions" developed by Egozcue and Díaz-Barrero (2003); Egozcue et al. (2006) and van den Boogaart (2005). The main goal is to compare the distribution of a given property in batches of grains (or crystals) from several sediment bodies (or rocks), and infer relations of evolution, proximity, distance, etc., between the sediment bodies (or rocks). As one deduces an "electrofacies" or a "petrofacies" from the electric-logged characteristics or the petrographic composition of a sediment/rock, we propose to define some sort of "densofacies" of the studied property as a descriptor of that sediment/rock

This is done in the following steps:

(1). choice of a (discretized) support for the desired property;

(2). estimation though kernel techniques of its density on that domain;

(3). division of the density by an adequate, scaled reference density (normal, exponential or uniform);

(4). regression of that relative density in log scale (completely removing nodes with estimated zero density) against some polynomials orthonormal with respect to the reference density (respectively Hermite, Laguerre or Legendre polynomials). This regression may be weighted using weights proportional to the reference density chosen.

The obtained regression coefficients (removing the coefficient-intercept) are adequately scaled with respect to each other, and characterize the density in a compact way, thus they can be used as input for further statistical treatment (covariance-based or distance-based). In the language of Hilbert spaces, these coefficients are *estimates* of the coordinates of the studied density with respect to the basis integrated by the (exponentiated) polynomials used.

There are many properties of single grains/crystals that can be treated in this way, though the most commonly measured one is its size/volume: this gives granulometric curves, one of the most informative characteristics of a sediment when studying its originating erosion, transport and deposition processes. For this particular case, an interesting application presented here is the compatibilization (a sort of "averaging") of grain size curves of the same sediment obtained with different measuring techniques. As a side result, we have also found that the "best" 2 parameters of a normal distribution in this framework are linearly related to the natural parameters of the corresponding exponential family (μ/σ^2 and $(1/\sigma^2 - 1)/\sqrt{2}$).

5 Acknowledgements

This research was funded by the the German Research Foundation (DFG, grant EY23/11-1) and the Department of Universities, Research and Information Society (DURSI, grant 2005 BP-A 10116) of the *Generalitat de Catalunya*. The authors are also very grateful to the fruitful discussions with J.J. Egozcue and V. Pawlowsky-Glahn.

REFERENCES

Barnes, S. and P. Roeder (2001). The range of spinel compositions in terrestrial mafic and ultramafic rocks. *Journal of Petrology 42*(12), 2279–2302.

van den Boogaart, K. (2005). Statistics structured by the Aitchison Space. In G. Mateu-Figueras and C. Barceló-Vidal (Eds.), *Compositional Data Analysis Workshop – CoDaWork'05, Proceedings.* Universitat de Girona, ISBN 84-8458-222-1, http://ima.udg.es/Activitats/CoDaWork05/.

van den Boogaart, K. G. and R. Tolosana-Delgado (2008). "compositions": a unified R package to analyze Compositional Data. *Computers and Geosciences 34*(4), 320–338.

Dick, H. and T. Bullen (1984). Chromian spinel as a petrogenetic indicator in abyssal and alpine-type peridotites and spatially associated lavas. *Contributions to Mineralogy and Petrology 86*(1), 54–76.

Egozcue, J. J. and J. L. Díaz-Barrero (2003). Hilbert space on probability density functions with Aitchison geometry. In S. Thió-Henestrosa and J. A. Martín-Fernández (Eds.), *Compositional Data Analysis Workshop – CoDaWork'03, Proceedings.* Universitat de Girona, ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork03/.

Egozcue, J. J., J. L. Díaz-Barrero, and V. Pawlowsky-Glahn (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica (English Series) 22*(4), 1175–1182. DOI: 10.1007/s10114-005-0678-2.

Kamenetsky, V., A. Crawford, and S. Meffre (2001). Factors controlling chemistry of magmatic spinel; an empirical study of associated olivine, cr-spinel and melt inclusions from primitive rocks. *Journal of Petrology 42*(4), 655–671.

Lemke, W. (1997). The Darss Sill in the southwestern Baltic sea – hydrographic and geological setting. See Pawlowsky-Glahn (1997), pp. 135–138.

Lwin, T. (2003). Parameterization of particle size distributions by three methods. *Mathematical Geology 35*(6), 719–736.

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1997). Different Classifications of the Darss Sill Data Set Based on Mixture Models for Compositional Data. See Pawlowsky-Glahn (1997), pp. 151–156.

McLaren, P. and D. Bowles (1985). The effects of sediment transport on grain-size distributions. *Journal of Sedimentary Petrology 55*(4), 457–470.

Pawlowsky-Glahn, V. (Ed.) (1997). *Proceedings of IAMG'97 — The third annual conference of the International Association for Mathematical Geology*, Volume I, II and addendum. International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E), 1100 p.

Sircombe, K. N. (2000). Quantitative comparison of large sets of geochronological data using multivariate analysis: A provenance study example from Australia. *Geochimica et Cosmochimica Acta 64*(9), 1593–1616.

Tauber, F. (1997). Treating grain-size data as continuous functions. See Pawlowsky-Glahn (1997), pp. 169–174.

Tauber, F. (1999). Spurious clusters in granulometric data caused by logratio transformation. *Mathematical Geology 31*(5), 491–504.