Stochastic Environmental Research and Risk Assessment 15 (2001) 384–398 © Springer-Verlag 2001

Geometric approach to statistical analysis on the simplex

V. Pawlowsky-Glahn, J. J. Egozcue

Abstract. The geometric interpretation of the expected value and the variance in real Euclidean space is used as a starting point to introduce metric counterparts on an arbitrary finite dimensional Hilbert space. This approach allows us to define general reasonable properties for estimators of parameters, like metric unbiasedness and minimum metric variance, resulting in a useful tool to better understand the logratio approach to the statistical analysis of compositional data, who's natural sample space is the simplex.

Key words: Aitchison geometry, compositional data, Euclidean space, finite dimensional Hilbert space, metric center, metric variance.

1 Introduction

The logratio approach to the statistical analysis of compositional data proposed by Aitchison (1982) has been the source of many discussions over the last decades. This is due to the enormous importance compositional data have in practice, as measurements in proportions of some whole, like percentages, ppm, etc. are extremely common in applied sciences. This approach makes it possible to perform classical statistical analysis on transformed data and to back transform the results, which is a clear advantage due to the large amount of methods available for multivariate normally distributed phenomena and the robustness of those. But there has been a certain reluctance in using the new approach by practitioners, which, besides the usual resistance to new theories, is due to the lack of classical properties of backtransformed estimators and models, like unbiasedness and minimum variance.

V. Pawlowsky-Glahn (✉)
Dept. d'Informàtica i Matemàtica Aplicada,
Universitat de Girona,
Campus Montilivi – P-1, E-17071 Girona, Spain
e-mail: vera.pawlowsky@udg.es

J. J. Egozcue
Dept. de Matemàtica Aplicada III, ETSECCPB,
Universitat Politècnica de Catalunya, Barcelona, Spain

The authors want to thank H. Burger and another anonymous referee for their comments, which helped to greatly improve the paper.

This research has been supported by the Dirección General de Enseñanza Superior (DGES) of the Spanish Ministry for Education and Culture through the project PB96-0501-C02-01.

In a recent paper, we have given a partial answer to these problems, based on concepts of metric center and metric variance related to the geometric structure of the simplex (Pawlowsky-Glahn and Egozcue, 2001). Here it is shown that the concepts of metric center and metric variance make sense for random vectors with sample space an arbitrary finite dimensional real Hilbert space. Using this approach, it is easy to proof essential properties for statistical inference not only on the simplex, which is the natural sample space for compositional data, but also in other sample spaces. Obviously, the same reasoning can be applied to complex spaces and there is no need to constrain it to elementary concepts and properties. But precisely elementary concepts and properties are useful to convince of the appropriateness and naturality of the definitions, showing that interpretation of real phenomena are much easier if we work on an appropriate sample space using the appropriate measures of central tendency and variability.

Throughout this work, we use the term finite dimensional real Hilbert space, instead of Euclidean space, for spaces with the appropriate structure that are different from m-dimensional real space $\mathbb{R}^m$. Although mathematically equivalent, we think that speaking about an Euclidean space, whether we refer to $\mathbb{R}^m$, or to its positive orthant $\mathbb{R}^m_+$, or to the interval (0,1), or to the simplex $\mathcal{S}^d_c$, can be easily misleading in this presentation.

The rationale behind the definitions and properties is related to that of Fréchet (1948) in his paper on random elements of arbitrary nature in a metric space. But Fréchet was primarily concerned with general, non-numerical spaces, while our interest lies in subsets of $\mathbb{R}^m$ with an appropriate structure. Given his approach, Fréchet was naturally interested in probabilistic problems, whereas we emphasize the estimation of parameters.

To illustrate our procedere, let us start recalling basic definitions and properties related to random variables in real space: given a continuous random variable X, the center or expected value is introduced as $\mathrm{E}[X] = \int_{-\infty}^{+\infty} x \, \mathrm{d}F_X(x)$, where $F_X(x)$ stands for the distribution function of X, and the variance as $\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]$. The geometric interpretation of these concepts is well known, and is often given either as a motivation or as an illustration. Nevertheless, the center can be defined as that value μ which minimizes the expected squared Euclidean distance $\mathrm{E}[d_e(X,\xi)^2]$, and the variance σ^2 can be defined as the expected value of the squared Euclidean distance around μ, $\sigma^2 = \mathrm{E}[d_e(X,\mu)^2]$. Obviously, $\mu = \mathrm{E}[X]$ and $\sigma^2 = \mathrm{Var}[X]$. To our understanding, this geometric approach gives its real meaning to the center as a measure of central tendency and to the variance as a measure of dispersion. Fréchet (1948) uses this philosophy to introduce the center and variance of a random vector with support an arbitrary metric space which is not necessarily a vector space. A similar reasoning lead Aitchison (2001) to justify the closed geometric mean as the natural center of a random composition and, later, Pawlowsky-Glahn and Egozcue (2001) to define the concept of metric variance for a random composition. Here, we extend this approach first to an arbitrary finite dimensional real Hilbert space, then we give some simple examples to illustrate its general interest, and finally we particularize on the simplex.

2
Notation and basic concepts

Given an m-dimensional real Hilbert space $\mathcal{E}$ (m-Hilbert space for short), with internal operation $\oplus$, external operation $\otimes$, and inner product $\langle .,. \rangle$, denote the associated norm by $\| \cdot \|$ and the associated distance by $d(.,.)$. We will use $\ominus$ and $\oslash$ whenever needed for the corresponding operations on the inverses and denote by $\mathbf{e}$ the neutral element with respect to the internal operation $\oplus$. This notation

$$\begin{tikzcd}
\Omega \arrow[r, "\mathbf{X}"] \arrow[dr, "h(\mathbf{X})"'] & \mathcal{E} \arrow[r, "d(\mathbf{X},\xi)"] \arrow[d, "h"'] & \mathbb{R} \\
 & \mathbb{R}^m \arrow[ur, "d_e(h(\mathbf{X}),h(\xi))"'] &
\end{tikzcd}$$

Fig. 1. Schematic representation of the relationship of a random vector with sample space ε and a random vector with sample space $\mathbb{R}^m$ through an isometric transformation h

has been chosen in order to relate and, at the same time, distinguish, operations in $\mathscr{E}$ from their counterparts in $\mathbb{R}^m$. For convenience, we will identify the scalar field with $(\mathbb{R}, +, \cdot)$. Recall that, in an m-Hilbert space, the distance is invariant with respect to the internal operation, as well as to the external operation,

$$d(\mathbf{x} \oplus \mathbf{z}, \mathbf{y} \oplus \mathbf{z}) = d(\mathbf{x}, \mathbf{y}); \quad d(\alpha \otimes \mathbf{x}, \alpha \otimes \mathbf{y}) = |\alpha| \cdot d(\mathbf{x}, \mathbf{y}) \ , \tag{1}$$

and that an m-Hilbert space is always isometric to a real Euclidean space, both having the same dimension m. This property is essential for subsequent developments. Thus, if we denote by h such an isometry, h is an isomorphism such that for any $\mathbf{x}, \mathbf{y} \in \mathscr{E}$,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \langle h(\mathbf{x}), h(\mathbf{y}) \rangle_e; \quad d(\mathbf{x}, \mathbf{y}) = d_e(h(\mathbf{x}), h(\mathbf{y})) \ ,$$

where $\langle .,. \rangle_e$ stands for the Euclidean inner product and $d_e(.,.)$ for the Euclidean distance in $\mathbb{R}^m$. A detailed account of these and related properties can be found in (Berberian, 1961).

In the sequel, following a standard approach as i.e. in (Ash, 1972), we consider random vectors X, defined as measurable functions from a probability space (Ω, $\mathscr{F}$, P_Ω) onto a sample space $\mathscr{E}$. Here Ω denotes an arbitrary set, $\mathscr{F}$ a σ-field of subsets of Ω, and P_Ω a probability measure on $\mathscr{F}$. A consistent definition of a random vector X requires a σ-field of subsets of $\mathscr{E}$. An easy and natural way to define it consists in taking $h^{-1}(\mathscr{B}(\mathbb{R}^m))$, being $\mathscr{B}(\mathbb{R}^m)$ the class of Borel sets of $\mathbb{R}^m$ (see Fig. 1 for a schematic illustration). With this definition $h(\mathbf{X})$ is a measurable function (i.e., a random vector) that goes from (Ω, $\mathscr{F}$, P_Ω) to $\mathbb{R}^m$. $\mathbf{X}$ induces a probability measure P on the σ-field $h^{-1}(\mathscr{B}(\mathbb{R}^m))$ of $\mathscr{E}$ and $h(\mathbf{X})$ induces a probability measure P_e on the σ-field $\mathscr{B}(\mathbb{R}^m)$ of $\mathbb{R}^m$. Note that for any set $A \in h^{-1}(\mathscr{B}(\mathbb{R}^m))$ we have $A = h^{-1}(B)$, for some $B \in \mathscr{B}(\mathbb{R}^m)$. Thus,

$$\begin{aligned} P[A] &= P[h^{-1}(B)] = P[\{\omega | \mathbf{X}(\omega) \in h^{-1}(B)\}] = P[\{\omega | h(\mathbf{X}(\omega)) \in B\}] \\ &= P_e[B] = P_e[h(A)] \ . \end{aligned}$$

With these probability measures we can define expectation in both spaces,

$$\mathrm{E}[\mathbf{X}] = \int\limits_{x \in \mathscr{E}} x \, \mathrm{d}P, \quad \mathrm{E}_e[h(\mathbf{X})] = \int\limits_{h(x) \in \mathbb{R}^m} h(x) \mathrm{d}P_e \ ,$$

satisfying $h(\mathrm{E}[\mathbf{X}]) = \mathrm{E}_e[h(\mathbf{X})]$. From now on, we will use the symbol $\mathrm{E}[\cdot]$ for both expectations. This concept of expectation is extended to functions g of $\mathbf{X} \in \mathscr{E}$ and g_e of $h(\mathbf{X}) \in \mathbb{R}^m$ as usual, resulting in

$$\mathrm{E}[g(\mathbf{X})] = \int\limits_{x \in \mathscr{E}} g(x)\mathrm{d}P, \quad \mathrm{E}_e[g_e(h(\mathbf{X}))] = \int\limits_{h(x) \in \mathbb{R}^m} g_e(h(x))\mathrm{d}P_e \ .$$

In particular, let $\xi \in \mathscr{E}$ be a fixed element and consider the function $d^2(\mathbf{X}, \xi)$, where $d(\ldots,\ldots)$ denotes the distance in $\mathscr{E}$ (Fig. 1). The expectation of such a function is then

$$\begin{aligned} \mathrm{E}[d^2(\mathbf{X}, \xi)] &= \int\limits_{x \in \mathscr{E}} d^2(x, \xi)\mathrm{d}P \\ &= \int\limits_{h(x) \in \mathbb{R}^m} d_e^2(h(x), h(\xi))\mathrm{d}P_e = \mathrm{E}[d_e^2(h(\mathbf{X}), h(\xi))] \ . \end{aligned} \tag{2}$$

Note that the latter expectation can also be defined using the corresponding probability measure induced in $\mathbb{R}_+$. In fact, $d^2(\mathbf{X}, \xi)$ is a univariate random variable with sample space $\mathbb{R}_+$ and its probability can be described by a univariate distribution function.

Now, following the rationale described in the introduction, let us introduce metric counterparts in $\mathscr{E}$ to usual measures of dispersion and central tendency in real Euclidean space.

3
Metric center and metric variance

Definition 1 The dispersion or metric variance around $\xi \in \mathscr{E}$ is the expected value of the squared distance between $\mathbf{X}$ and ξ: $\mathrm{Mvar}[\mathbf{X}, \xi] = \mathrm{E}[d^2(\mathbf{X}, \xi)]$, provided that the last expectation exists.

Note that the metric variance is well defined, given that the squared distance is a real function. Assuming the metric variance of $\mathbf{X}$ exists, we can introduce now the metric center of its distribution as follows.

Definition 2 The metric center of the distribution of $\mathbf{X}$ is that element $\xi \in \mathscr{E}$ which minimizes $\mathrm{Mvar}[\mathbf{X}, \xi]$. It is called metric center of $\mathbf{X}$ and is denoted by $\mathrm{Mcen}[\mathbf{X}]$ for short.

Following our strategy to paraphrase standard statistical concepts, to call metric variance the metric variance around $\mathrm{Mcen}[\mathbf{X}]$ and metric standard deviation its square root is only natural. We state this as a definition for easy of reference.

Definition 3 The metric variance around the metric center $\mathrm{Mcen}[\mathbf{X}]$ of the distribution of $\mathbf{X}$ is given by $\mathrm{Mvar}[\mathbf{X}, \mathrm{Mcen}[\mathbf{X}]] = \mathrm{E}[d^2(\mathbf{X}, \mathrm{Mcen}[\mathbf{X}])]$. It is called metric variance and is denoted by $\mathrm{Mvar}[\mathbf{X}]$ for short. The square root of the metric variance of a random composition is called metric standard deviation and is denoted by $\mathrm{Mstd}[\mathbf{X}]$.

Given the existence of an isometry h between the m-Hilbert space $\mathscr{E}$ and real m-Euclidean space, it is clear that we can transfer directly properties derived from the geometric structure between them. To do so, the following two propositions are essential.

Proposition 1: *If* $h : \mathscr{E} \rightarrow \mathbb{R}^m$ *is an isometry, then* $\mathrm{Mcen}[\mathbf{X}] = h^{-1}(\mathrm{E}[h(\mathbf{X})])$.

Proof: If $h : \mathscr{E} \rightarrow \mathbb{R}^m$ is an isometry, then it holds that $\mathrm{E}[d^2(\mathbf{X}, \xi)] = \mathrm{E}[d_e^2(h(\mathbf{X}), h(\xi))]$ (see Eq. (2)), and the vector that minimizes $\mathrm{E}[d_e^2(h(\mathbf{X}), h(\xi))]$ is $h(\xi) = \mathrm{E}[h(\mathbf{X})]$. Consequently, the vector that minimizes $\mathrm{E}[d^2(\mathbf{X}, \xi)]$ is $\xi = h^{-1}(\mathrm{E}[h(\mathbf{X})])$. □

Proposition 2: *If* $h : \mathscr{E} \rightarrow \mathbb{R}^m$ *is an isometry and* $h(\mathbf{X}) = \mathbf{Y} \in \mathbb{R}^m$, *then*

$$\mathrm{Mvar}[\mathbf{X}] = \sum_{i=1}^{m} \mathrm{Var}[Y_i] \ .$$

Proof: Being h an isometry and taking into account Proposition 1,

$$\mathrm{Mvar}[\mathbf{X}] = \mathrm{E}[d^2(\mathbf{X}, \mathrm{Mcen}[\mathbf{X}])] = \mathrm{E}[d_e^2(h(\mathbf{X}), \mathrm{E}[h(\mathbf{X})])]$$

holds. Writing $h(\mathbf{X}) = \mathbf{Y}$ and using the definition of Euclidean distance the equality is obtained. □

Note the assumption h is an isometry is actually too strong for Proposition 1 to hold, although it simplifies the proof. In fact, if h is an isomorphism, Proposition 1 holds too. Nevertheless, the isometry assumption is necessary for Proposition 2 to hold.

Given the linearity of h and $\mathrm{E}[\cdot]$ in Propositions 1 and 2, classical properties of center and variance in real Euclidean space related to linearity and translation invariance, as well as unicity, are transfered to the metric center and metric variance in $\mathscr{E}$. In particular, denoting by $\mathbf{X}$, $\mathbf{Y}, \ldots$ random vectors with sample space $\mathscr{E}$ and assuming that the expectations involved exist, the following properties, which are stated without proof, hold:

Proposition 3: *For all* $\mathbf{b} \in \mathscr{E}$ *and all* $\alpha \in \mathbb{R}$,

$$\mathrm{Mcen}[(\alpha \otimes \mathbf{X}) \oplus \mathbf{b}] = (\alpha \otimes \mathrm{Mcen}[\mathbf{X}]) \oplus \mathbf{b} \ .$$

Proposition 4: $\mathrm{Mcen}[\mathbf{X} \ominus \mathrm{Mcen}[\mathbf{X}]] = \mathbf{e}$.

Proposition 5: $\mathrm{Mcen}[\mathbf{X} \oplus \mathbf{Y}] = \mathrm{Mcen}[\mathbf{X}] \oplus \mathrm{Mcen}[\mathbf{Y}]$, *and, in general,*

$$\mathrm{Mcen}[\mathbf{X}_1 \oplus \mathbf{X}_2 \oplus \cdots \oplus \mathbf{X}_N] = \mathrm{Mcen}[\mathbf{X}_1] \oplus \mathrm{Mcen}[\mathbf{X}_2] \oplus \cdots \oplus \mathrm{Mcen}[\mathbf{X}_N] \ .$$

For simplicity, in what follows we will use the notation $\bigoplus_{n=1}^{N} \mathbf{X}_n = \mathbf{X}_1 \oplus \mathbf{X}_2 \oplus \cdots \oplus \mathbf{X}_N$.

Proposition 6: *For all* $\mathbf{b} \in \mathscr{E}$ *and all* $\alpha \in \mathbb{R}$,

$$\mathrm{Mvar}[(\alpha \otimes \mathbf{X}) \oplus \mathbf{b}] = \alpha^2 \mathrm{Mvar}[\mathbf{X}] \ .$$

Proposition 7: *For all* $\mathbf{b} \in \mathscr{E}$

$$\mathrm{Mvar}[\mathbf{X}] = \mathrm{Mvar}[\mathbf{X}, \mathbf{b}] - d^2(\mathrm{Mcen}[\mathbf{X}], \mathbf{b}) \ .$$

Remark 1 Setting in the last proposition $\mathbf{b} = \mathbf{e}$, the result is analogous to the standard relationship for univariate real random variables, namely

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 = \mathrm{E}[(X-0)^2] - (\mathrm{E}[X]-0)^2 \\ &= \mathrm{E}[d_e^2(X, 0)] - d_e^2(\mathrm{E}[X], 0) \ .\end{aligned}$$

In what follows, it is understood that independence between random vectors which sample space is $\mathscr{E}$ is defined in an analogous manner to the standard way (Ash, 1972: p. 213).

Proposition 8: *For* $\mathbf{X}$, $\mathbf{Y}$ *independent*,

$$\mathrm{Mvar}[\mathbf{X} \oplus \mathbf{Y}] = \mathrm{Mvar}[\mathbf{X}] + \mathrm{Mvar}[\mathbf{Y}] \ ,$$

and, in general, for $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ *jointly independent*,

$$\mathrm{Mvar}\left[\bigoplus_{n=1}^{N} \mathbf{X}_n\right] = \sum_{n=1}^{N} \mathrm{Mvar}[\mathbf{X}_n] \ .$$

A simple property, but of primary importance, is the Chebyshev inequality that holds for the metric center and metric variance. It gives us a further interpretation of the metric variance, or its squared root, as a dispersion measure around the metric center, independently of the distribution of the random vector.

Proposition 9: Chebyshev inequality. *For any* $k > 0$,

$$\mathrm{P}[d(\mathbf{X}, \mathrm{Mcen}[\mathbf{X}]) \geq k\mathrm{Mstd}[\mathbf{X}]] \leq \frac{1}{k^2} \ .$$

Proof: The standard proof of the Chebyshev inequality follows. Define the set

$$A = \{\mathbf{x} \in \mathscr{E} : d(\mathbf{x}, \mathrm{Mcen}[\mathbf{X}]) \geq k\mathrm{Mstd}[\mathbf{X}]\} \ .$$

Then,

$$\mathrm{Mvar}[\mathbf{X}] \geq \mathrm{E}[d^2(\mathbf{X}, \mathrm{Mcen}[\mathbf{X}])|A] \geq k^2 \mathrm{Mvar}[X] \ P[\mathbf{X} \in A] \ ,$$

from which the statement holds. □

Thus, the metric variance, defined as the expected value of the distance to the metric center, allows us a geometric understanding of the measures of central tendency and dispersion of a random vector in a given m-Hilbert space.

Before we proceed, it is worthwhile to note that Rao (1982) introduced the concept of quadratic entropy, which is equivalent to our metric variance, and that

Cuadras et al. (1997) introduced equivalent concepts with the purpose of classification. The difference is that they do not insist explicitly in the distinction between random vectors which sample space is not real Euclidean space, specially when it comes to estimation problems. The importance of this distinction will be seen in the next section.

4 Estimation

Consider a random vector $\mathbf{X}$ which sample space is $\mathscr{E}$, and a random sample of size N, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$. Assume that the probability measure of $\mathbf{X}$, $P_{\mathscr{E}}$, depends on an unknown parameter, or vector of parameters, $\theta \in \Theta$, where Θ stands for the parameter space. Θ is also assumed to have an m_θ-dimensional real Hilbert space structure. The purpose of this section is to define basic desirable properties of estimators $\hat{\theta}$ of θ. Since these definitions, being parallel to ordinary ones, depend on the metric defined in Θ, we introduce an 'M$_\theta$' preceeding the concept in order to distinguish them; the subindex θ is introduced to make clear that the parameter space and the sample space are not necessarily the same. Therefore, whenever we refer to the metric in the sample space $\mathscr{E}$, we will use the notation 'M$_{\mathscr{E}}$'. Although tedious, we insist in this notation to state clearly that the essential idea of metric properties is to consider them in the appropriate space. We also point out that we are going to use $\mathrm{Mcen}_\theta[\hat{\theta}(\mathbf{X}_1, \ldots, \mathbf{X}_N)]$, where $\hat{\theta}$ is a function of the random sample. This metric center implicitly requires the definition of expectation and the corresponding probability measure. If P is the probability measure induced by $\mathbf{X}$ in $\mathscr{E}$, the joint probability measure of the random sample is obtained as the direct product of P as many times as N. The expectation is then taken as an integral of $\hat{\theta}$ with respect to this probability measure. Note that the subscript in $\mathrm{Mcen}_\theta[\hat{\theta}]$ is related to Θ, the sample space of the function $\hat{\theta}$, and not to the definition of the probability measure.

At this point, the way of reasoning may differ from standard approaches because the m_θ-Hilbert space structure of Θ is normally not assumed to be different from $\mathbb{R}^m$, being $\Theta \subset \mathbb{R}^m$. Whenever Θ can be identified with $\mathbb{R}^k$ for some integer k, the following approach coincides with the standard approach. But, if Θ is not a linear subspace of $\mathbb{R}^m$, then the present approach claims for a new m_θ-Hilbert space structure in Θ and the equivalence no longer holds.

Definition 4 $\hat{\theta}$ is an M$_\theta$-centered or M$_\theta$-unbiased estimator of θ if, and only if, the metric center of $\hat{\theta}$, with respect to the metric defined in the parameter space Θ, is the unknown parameter θ: $\mathrm{Mcen}_\theta[\hat{\theta}] = \theta$.

Using Proposition 4 we obtain the equivalent property:

Proposition 10: $\mathrm{Mcen}_\theta[\hat{\theta}] = \theta$ *is equivalent to* $\mathrm{Mcen}_\theta[\hat{\theta} \oplus_\theta \theta^{-1}] = \mathbf{e}_\theta$, *the neutral element of the internal operation* $\oplus_\theta$ *on* Θ.

Definition 5 $\mathrm{Mcen}_\theta[\hat{\theta} \oplus_\theta \theta^{-1}]$ is called the M$_\theta$-bias of $\hat{\theta}$.

Using now Proposition 3 we obtain:

Proposition 11: $\mathrm{Mcen}_\theta[\hat{\theta} \oplus_\theta \theta^{-1}] = \mathrm{Mcen}_\theta[\hat{\theta}] \oplus_\theta \theta^{-1} = \mathrm{Mcen}_\theta[\hat{\theta}] \ominus_\theta \theta$.

In order to compare the M$_\theta$-bias from different estimators of the same parameter θ, the distance to $\mathbf{e}_\theta$ can be used, as they belong to the same space. Thus, the adequate measure to be used is

$$d_\theta(\mathrm{Mcen}_\theta[\hat{\theta}] \oplus_\theta \theta^{-1}, \mathbf{e}_\theta) = d_\theta(\mathrm{Mcen}_\theta[\hat{\theta}], \theta) \ \ ,$$

where the equality is derived from Eq. (1).

The mean quadratic error is another important criterion in estimation of standard parameters. The analogous in our case is the following.

Definition 6 $\mathrm{Mvar}_\theta[\hat{\theta}, \theta]$ is called the M$_\theta$-quadratic error of $\hat{\theta}$.

This definition of M$_\theta$-quadratic error has similar properties to standard quadratic error. Particularly, applying property 7, it is related with the M$_\theta$-bias and the metric variance of the estimator in the same way the standard estimators are: quadratic error equals squared bias plus variance of the estimator.

Proposition 12: $\mathrm{Mvar}_\theta[\hat{\theta}, \theta] = \mathrm{Mvar}_\theta[\hat{\theta}] + d_\theta^2(\mathrm{Mcen}_\theta[\hat{\theta}], \theta)$.

After these definitions, general concepts on estimation of standard parameters can be easily extended to metric counterparts (e.g. asymptotically unbiased estimators, consistency in mean quadratic error). In the present context we are specially interested in the following definitions, related to the so called best linear unbiased estimators (BLUE).

Definition 7 Given two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of $\theta \in \Theta$, $\hat{\theta}_1$ is said to be more M$_\theta$-efficient than $\hat{\theta}_2$ with respect to the distance defined in Θ if, and only if, $\mathrm{Mvar}_\theta[\hat{\theta}_1, \theta] < \mathrm{Mvar}_\theta[\hat{\theta}_2, \theta]$.

Definition 8 Given a class $\hat{\Theta} \subset \Theta$ of estimators of θ, $\hat{\theta} \in \hat{\Theta}$ is said to be M$_\theta$-best within the class $\hat{\Theta}$ if, and only if, it is M$_\theta$-centered and $\mathrm{Mvar}_\theta[\hat{\theta}] < \mathrm{Mvar}_\theta[\hat{\theta}_i]$ for all $\hat{\theta}_i \in \hat{\Theta}$; i.e. it is the most M$_\theta$-efficient among the M$_\theta$-centered estimators in $\hat{\Theta}$.

Obviously, other standard characterizations of estimators, usual in the context of random variables with support the real line, can be given, simply by substituting the Euclidean distance by the appropriate distance defined in the parameter space, and the expected value by the M$_\theta$-center, but it goes beyond the purpose of this paper. Therefore, let us proceed to define an M$_\theta$-best linear estimator of a parameter θ within the class of linear estimators of θ, where linear is understood in the following sense:

Definition 9 Given a function $g(\cdot)$ from $\mathscr{E}$ onto Θ,

$$\hat{\theta} = \bigoplus_{n=1}^{N} (\alpha_n \otimes_\theta g(\mathbf{X}_n)) \ \ ,$$

is said to be an M$_\theta$-linear g-function of the sample $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$. Whenever $g(\cdot)$ is suitable for estimation of θ, then $\hat{\theta}$ is said to be an M$_\theta$-linear g-estimator of θ.

Now, for $g(\cdot)$ a function from $\mathscr{E}$ onto Θ, the following propositions can be set forth.

Proposition 13: *If, for any $n = 1, \ldots, N$, $\mathrm{Mcen}_\theta[g(\mathbf{X}_n)] = \theta$, then*

$$\hat{\theta} = \bigoplus_{n=1}^{N} \left(\frac{1}{N} \otimes_\theta g(\mathbf{X}_n) \right)$$

is an M$_\theta$*-linear and* M$_\theta$*-unbiased g-estimator of θ.*

Proof: $\hat{\theta}$ is an M_θ-linear function by definition. Taking metric centers in the definition of the g-estimator and using Propositions 5 and 3, the fact that $\mathrm{Mcen}_\theta[g(\mathbf{X}_n)] = \theta$, and standard properties of vector spaces, the statement holds. Note that, as usual, independence of the sample is not a requirement for the proof. □

Proposition 14: *Given a function* $g : \mathscr{E} \to \Theta$ *such that* $\mathrm{Mcen}_\theta[g(\mathbf{X}_n)] = \theta$,

$$\hat{\theta} = \bigoplus_{n=1}^{N} \left(\frac{1}{N} \otimes_\theta g(\mathbf{X}_n) \right)$$

is the M_θ*-best g-estimator of* θ *within the class of* M_θ*-linear* M_θ*-unbiased g-estimators of* θ*. Moreover,*

$$\mathrm{Mvar}_\theta[\hat{\theta}] = \frac{\mathrm{Mvar}_\theta[g(\mathbf{X})]}{N} \ .$$

Proof: Consider a general M_θ-linear M_θ-unbiased g-estimator of θ

$$\tilde{\theta} = \bigoplus_{n=1}^{N} (\alpha_n \otimes_\theta g(\mathbf{X}_n)) \ .$$

If $\tilde{\theta}$ is M_θ-unbiased, then, using again Propositions 5 and 3 as well as standard operations in a vector space, we obtain $\sum_{n=1}^{N} \alpha_n = 1$. To see that it is M_θ-best we have to see that the metric variance reaches a minimum when $\alpha_n = 1/N$. Given that by definition of random sample we have $\mathrm{Mvar}_\theta[g(\mathbf{X}_n)] = \mathrm{Mvar}_\theta[g(\mathbf{X})]$ and independence of the sample, the metric variance of θ can be expressed, using Propositions 8 and 6, as

$$\mathrm{Mvar}_\theta[\tilde{\theta}] = \mathrm{Mvar}_\theta[g(\mathbf{X})] \left(\sum_{n=1}^{N} \alpha_n^2 \right) \ ,$$

which is minimum when, for $n = 1, \ldots, N$, $\alpha_n = 1/N$. Therefore, for minimum metric variance $\tilde{\theta} = \hat{\theta}$ and $\mathrm{Mvar}_\theta[\hat{\theta}] = \mathrm{Mvar}_\theta[g(\mathbf{X})]/N$. □

In Proposition 14 the M_θ-linear, M_θ-unbiased, M_θ-best g-estimator has been identified, but the function $g : \mathscr{E} \to \Theta$ was beforehand given. A natural extension may be to make g free within a given class of functions. Although a detailed discussion of such a case is out of the scope of this presentation, the proof of Proposition 14 points out that minimization of $\mathrm{Mvar}_\theta[\hat{\theta}]$ can be decomposed into two steps: search for the best $g : \mathscr{E} \to \Theta$ and optimization for the α_n's; and, then, the optimum value of α_n is still $1/N$.

Proposition 15: *Given a function* $g : \mathscr{E} \to \Theta$ *such that* $\mathrm{Mcen}_\theta[g(\mathbf{X}_n)] = \theta$, $\hat{\theta} = \bigoplus_{n=1}^{N} \left(\frac{1}{N} \otimes_\theta g(\mathbf{X}_n) \right)$ *satisfies the weak law of large numbers given by*

$$P\left[d_\theta(\hat{\theta}, \theta) \geq \frac{\mathrm{Mstd}_\theta[g(\mathbf{X})]}{\sqrt{\varepsilon N}} \right] \leq \varepsilon \ ,$$

for any $\varepsilon > 0$. *Consequently*, $\hat{\theta}$ *converges in probability to* $\mathrm{Mcen}_\theta[g(\mathbf{X})]$ *for* $N \to \infty$.

Proof: This standard result is obtained by applying the Chebyshev inequality stated in Proposition 9 to the random function $\hat{\theta}$, which metric center and metric variance are, respectively, θ and $\mathrm{Mvar}_\theta[g(\mathbf{X})]/N$. Setting $1/h^2 = \varepsilon$, the desired result is obtained. □

Propositions 13–15 clearly establish that $\hat{\theta} = \bigoplus_{n=1}^{N} \left(\frac{1}{N} \otimes_\theta g(\mathbf{X}_n)\right)$ is an M_θ-linear M_θ-unbiased g-estimator of θ, which is M_θ-best with respect to the corresponding distance defined in the parameter space Θ.

Example 1: Univariate random variables with sample space the real line.

For univariate random variables with sample space $\mathbb{R} = \mathscr{E}$ (i.e. support the real line), it is straightforward to obtain all the standard results. In fact, $\mathbb{R}$ is a real 1-Hilbert space with the usual operations: addition for the internal or Abelian group operation and product for the external operation. The inner product is the usual one (i.e. the product), and the distance is the Euclidean distance. The metric center is then nothing else but the usual expected value, and the best, linear, unbiased estimator associated to the Euclidean distance is the average or arithmetic mean of the sample.

Example 2: Univariate random variables with sample space the positive real line.

A particularly interesting case is that of an univariate random variable X, which sample space (i.e. support) is the positive real line $\mathbb{R}_+ = \mathscr{E}$. This sample space is a 1-Hilbert space with the following operations and definitions. Let $x, y \in \mathbb{R}_+$ and $\alpha \in \mathbb{R}$; then we have

1. Internal, Abelian group operation: $x \oplus_+ y = x \cdot y$.
2. External operation: $\alpha \otimes_+ x = x^\alpha$.
3. Inner product: $\langle x, y \rangle_+ = \ln x \cdot \ln y$.
4. Distance: $d_+(x, y) = |\ln x - \ln y|$.
5. Norm: $\|x\|_+ = |\ln x|$.
6. Isometry $h : \mathbb{R}_+ \to \mathbb{R}$ such that $h(x) = \ln x$, with inverse $h^{-1}(y) = \exp(y)$.

The metric center in $\mathbb{R}_+$ with this structure is given in Proposition 1 as

$$\gamma = \mathrm{Mcen}_+[X] = \exp(\mathrm{E}[\ln X]) \ .$$

γ is again a value in $\mathbb{R}_+$ and, therefore, the sample space of γ is $\mathbb{R}_+$, which 1-Hilbert space structure has been previously defined. Given a random sample $X_1, \ldots, X_N$, to estimate γ we can take in Definition 9 the function $g = \mathrm{id}$, the identity in $\mathbb{R}^+$. Then, for any n, $\mathrm{Mcen}_+[g(X_n)] = \mathrm{Mcen}_+[X] = \gamma$, and Propositions 13 and 14 state that $\hat{\gamma} = \Pi_{n=1}^{N} X_n^{1/N}$ is the M_+-best id-estimator of γ within the class of M_+-linear, M_+-unbiased id-estimators of γ. Note that the estimator obtained is the geometric mean of the sample and the estimated parameter is $\gamma = \exp(\mathrm{E}[\ln X])$, also known as the theoretical geometric mean of a random variable. These facts recall us the standard treatment of lognormal variates and state that, in terms of the geometric structure of $\mathbb{R}_+$, the natural measure of central tendency is the theoretical geometric mean and the best estimator is the geometric mean of the sample.

Note that this reasoning can be applied to any measure of difference, which sample space is by definition $\mathbb{R}_+$. As a result, we obtain an estimator for the metric variance which is different from the one obtained by the method of moments in real Euclidean space.

Example 3: Univariate random variable with sample space $I = (0,1)$.

Let X be an univariate random variable which sample space is $I = (0,1)$. This sample space is a 1-Hilbert space with the following operations and definitions. Let $x, y \in I$ and $\alpha \in \mathbb{R}$, then we have:

1. Internal, Abelian group operation:

$$x \oplus_I y = \frac{xy}{(1-x)(1-y)+xy} \quad .$$

2. External operation:

$$\alpha \otimes_I x = \frac{x^\alpha}{(1-x)^\alpha + x^\alpha} \quad .$$

3. Inner product:

$$\langle x, y \rangle_I = \ln \frac{x}{1-x} \cdot \ln \frac{y}{1-y} \quad .$$

4. Distance:

$$d_I(x,y) = \left| \ln \frac{x(1-y)}{y(1-x)} \right| \quad .$$

5. Norm:

$$\|x\|_I = \left| \ln \frac{x}{1-x} \right| \quad .$$

6. Isometry (logit transformation):

$$h : I \longrightarrow \mathbb{R} \text{ such that } h(x) = \ln \frac{x}{1-x}, \quad \text{with inverse } h^{-1}(y) = \frac{\exp(y)}{1+\exp(y)} \quad .$$

According to Proposition 1, the metric center in I is given by

$$\gamma = \mathrm{Mcen}_I[X] = \frac{\exp(\mathrm{E}[\ln(X/(1-X))])}{1+\exp(\mathrm{E}[\ln(X/(1-X))])} \quad ,$$

and γ is again a value in I. Thus, the parameter space of γ is I, which 1-Hilbert space structure has been just defined. Given a random sample $X_1, \ldots, X_N$, to estimate γ we can take in Definition 9 the function $g = \mathrm{id}$, the identity in I. Then, for any n, $\mathrm{Mcen}_I[g(X_n)] = \mathrm{Mcen}_I[X] = \gamma$, and Propositions 13 and 14 state that

$$\hat{\gamma} = \frac{\prod_{n=1}^N X_n^{1/N}}{\prod_{n=1}^N (1-X_n)^{1/N} + \prod_{n=1}^N X_n^{1/N}}$$

is the M_I-best id-estimator of γ within the class of M_I-linear, M_I-unbiased id-estimators of γ.

5
Estimation on the simplex

Recall that $\mathbf{x} = (x_1, \ldots, x_d)'$ is by definition a d-part composition if, and only if, all its components are strictly positive real numbers and their sum is a constant c. The constant c is 1 if measurements are made in parts per unit, or 100 if measurements are made in percent. The sample space of d-part compositional data with constant sum c is thus the simplex

$$\mathcal{S}_c^d = \left\{ \mathbf{x} = (x_1, \ldots, x_d)' | x_i > 0, i = 1, \ldots, d; \sum_{i=1}^{d} x_i = c \right\} ,$$

where the prime stands for transpose. Although mathematically less comfortable, we keep the constant c in the definition and in the notation, to avoid confusion arising from the fact that in geology it is more common to use $c = 100$ than $c = 1$. But, to simplify the mathematical developments, we include the constant in the closure operation as stated below.

Basic operations on the simplex have been introduced by Aitchison (1986). They are the perturbation operation, defined for any two vectors $\mathbf{x}, \mathbf{y} \in \mathcal{S}_c^d$ as

$$\mathbf{x} \circ \mathbf{y} = \mathcal{C}(x_1 y_1, \ldots, x_d y_d)' \ , \tag{3}$$

and the power transformation, defined for a vector $\mathbf{x} \in \mathcal{S}_c^d$ and a scalar $\alpha \in \mathbb{R}$ as

$$\alpha \diamond \mathbf{x} = \mathcal{C}(x_1^{\alpha}, \ldots, x_d^{\alpha})' \ , \tag{4}$$

where the $\mathcal{C}$ denotes the closure operation defined for a vector $\mathbf{z} = (z_1, \ldots, z_d)'$ as

$$\mathcal{C}(\mathbf{z}) = \mathcal{C} \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_d \end{pmatrix} = \begin{pmatrix} \frac{c \cdot z_1}{z_1 + z_2 + \cdots + z_d} \\ \frac{c \cdot z_2}{z_1 + z_2 + \cdots + z_d} \\ \vdots \\ \frac{c \cdot z_d}{z_1 + z_2 + \cdots + z_d} \end{pmatrix} .$$

Perturbation and power transformation induce a vector space structure in the simplex. Then, to obtain a $(d-1)$-Hilbert space structure on $\mathcal{S}_c^d$, the following inner product and associated norm and distance can be used (Aitchison, 2001; Pawlowsky-Glahn and Egozcue, 2001):

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{d} \sum_{i<j} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} ;$$

$$\|\mathbf{x}\|_a = \sqrt{\frac{1}{d} \sum_{i<j} \left(\ln \frac{x_i}{x_j} \right)^2} ;$$

$$d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{d} \sum_{i<j} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2} .$$

To refer to the properties of $(\mathcal{S}_c^d, \circ, \diamond)$ as a $(d-1)$-Hilbert space, we shall talk globally about the Aitchison geometry on the simplex, and in particular about the Aitchison distance, norm and inner product.

$\mathcal{S}_c^d$ is a $(d-1)$-Hilbert space and thus there exists an isometry between $\mathcal{S}_c^d$ and $\mathbb{R}^{d-1}$ which could be used to analize results from previous sections in the particular case of random compositions. Nevertheless, to simplify presentation, we will use the clr transformation defined by Aitchison (1986), which is an isometry between the simplex with the Aitchison geometry and the $(d-1)$-hyperplane going through the origin and parallel to the simplex in $\mathbb{R}^d$ and equipped with the usual Euclidean geometry in $\mathbb{R}^{d-1}$ projected from the real Euclidean space $\mathbb{R}^d$. Recall that the clr transformation is defined as

$$\mathrm{clr}(\mathbf{x}) = \left(\ln\frac{x_1}{g(\mathbf{x})}, \ln\frac{x_2}{g(\mathbf{x})}, \ldots, \ln\frac{x_d}{g(\mathbf{x})}\right) ,$$

where $g(\mathbf{x}) = \left(\prod_{i=1}^{d} x_i\right)^{1/d}$, the geometric mean of $\mathbf{x}$. The inverse is obtained by taking first exponentials and then applying the closure operation, which cancels out multiplicative constants.

With these elements at hand we can proceed to analyze the metric variance and metric center of the distribution of a random vector $\mathbf{X}$ with sample space $\mathcal{S}_c^d$.

Proposition 16: *The metric center* $\mathrm{Mcen}_a[\mathbf{X}]$ *of* $\mathbf{X}$ *is the center or closed geometric mean of* $\mathbf{X}$,

$$\mathrm{Mcen}_a[\mathbf{X}] = \mathscr{C}(\exp\{\mathrm{E}[\ln(X_1)]\}, \ldots, \exp\{\mathrm{E}[\ln(X_d)]\})' .$$

This result is actually the original definition of center of a random composition given by Aitchison (1997). It is obtained by applying Proposition 1 with the clr transformation.

Proposition 17: *The metric variance can be expressed as*,

$$\mathrm{Mvar}_a[\mathbf{X}] = \frac{1}{d}\sum_{i<j} \mathrm{Var}\left[\ln\frac{X_i}{X_j}\right] = \sum_{i=1}^{d} \mathrm{Var}\left[\ln\frac{X_i}{g(\mathbf{X})}\right] .$$

The first equality states that the metric variance with respect to the Aitchison distance is identical to the total variance defined by Aitchison (1997). It is derived directly from the definition of metric variance and of the Aitchison distance. The second equality is obtained using Proposition 2 and the clr transformation.

Note that Proposition 16 implies that, writing $\mathrm{Mcen}_a[\mathbf{X}] = \gamma$, for $i, j = 1, \ldots, d$,

$$\mathrm{E}\left[\ln\frac{X_i}{X_j}\right] = \ln\frac{\gamma_i}{\gamma_j} .$$

As a result of these statements, we can say that the closed geometric mean of random compositions minimizes the metric variance on the simplex with respect to the Aitchison distance. We can also say that the total variance defined by Aitchison (1997) is an appropriate measure of compositional variability within the simplex, as it coincides with the expected value of the squared Aitchison distance to the metric center of the distribution.

Looking at other properties of random compositions derived from propositions stated in Sect. 3, we see that perturbation of a random composition affects the metric center in that it leads to a perturbed metric center (Proposition 3), whereas it has no effect on the metric variance (Proposition 6). As a consequence, we can center the random composition by perturbing it with the inverse of the metric center (proposition 4), thus giving theoretical support to the approach presented in (Buccianti et al., 1999; Martín-Fernández et al., 1999; Eynatten et al., 2001). Furthermore, the metric center is linear with respect to perturbation (Proposition 5), whereas this property holds for the metric variance only in case of independence of the random compositions involved (Proposition 8).

Proposition 7 applied to random compositions on the simplex tells us, that the metric variance can be expressed as the metric variance around an arbitrary point **b** in the simplex minus the squared Aitchison distance between the metric center and the same point. Substituting **b** by the baricenter or neutral element of perturbation, **e**, gives the following result:

$$\mathrm{Mvar}_a[\mathbf{X}] = \mathrm{E}[d_a^2(\mathbf{X}, \mathbf{e})] - d_a^2(\mathrm{Mcen}_a[\mathbf{X}], \mathbf{e}) \ ,$$

suggesting an analogy to central and non-central moments in real space.

Another interesting feature is related to the power transformation. The power transformation of a random composition multiplies the metric center (Proposition 3) and its square multiplies the metric variance (Proposition 6). Thus, we can introduce an equivalent concept to standardized random vectors by using perturbation with the inverse of the metric center and power transformation with the inverse of the metric standard deviation to obtain random compositions centered at the baricenter **e** and with unit variance:

$$U = \frac{1}{\mathrm{Mstd}_a[\mathbf{X}]} \diamond \left(\mathbf{X} \circ (\mathrm{Mcen}_a[\mathbf{X}])^{-1}\right) \ .$$

Finally, the Chebyshev inequality stated in Proposition 9 gives us a way to obtain regions within the simplex where we have a probability smaller or equal to $1/k^2$ that the random composition is at a distance from the metric center larger then k times the metric standard deviation.

Concerning the estimation of the center, we can say that, taking in Proposition 14 the identity function $g(\mathbf{X}_n) = \mathrm{id}(\mathbf{X}_n) = \mathbf{X}_n$ we obtain that

$$\bar{\mathbf{X}}_a = \bigcirc_{n=1}^{N} \left(\frac{1}{N} \diamond \mathbf{X}_n\right)$$

is the M_a-best id-estimator of $\mathrm{Mcen}_a[\mathbf{X}]$ within the class of M_a-linear M_a-unbiased id-estimators of $\mathrm{Mcen}_a[\mathbf{X}]$. Moreover,

$$\mathrm{Mvar}_a[\bar{\mathbf{X}}_a] = \frac{\mathrm{Mvar}_a[\mathbf{X}]}{N} \ .$$

These are only the basic properties of the metric center, but it is clear that the same rationale would lead us to transfer whatsoever properties based on Euclidean reasoning from real space into the simplex. This approach assures us that we will obtain properties of optimality in the simplex, completely equivalent to those in real space.

6
Conclusions

The existence of an appropriate m-Hilbert space structure in the simplex suggests a different approach to the statistical analysis of compositional data based on geometric reasoning. Based on this approach, which is completely parallel to the usual one in Euclidean space, it is straightforward to define reasonable properties for estimators of compositional parameters. It assures us that we will obtain properties of optimality, completely equivalent to those in real space, in the simplex. In particular, the closed geometric mean is a linear, unbiased estimator that minimizes the metric variance with respect to the Aitchison geometry on the simplex.

But even more important is, that the same methodology allows us to study properties of probability measures on any sample space with an appropriate finite dimensional real Hilbert space structure, thus opening up a geometric approach to the study of statistical properties in general. Furthermore, it has been shown that this approach is also valid for estimators of unknown parameters of a probability measure and/or characteristics of a random vector, like the metric center. As a consequence, care has to be applied in analyzing the structure of the parameter space to assure the appropriateness of applied methods.

References

Aitchison J (1982) The statistical analysis of compositional data (with discussion). J. Royal Stat. Soc., Series B (Statistical Methodology) 44(2): 139–177

Aitchison J (1986) The Statistical Analysis of Compositional Data. Monographs on Statistics and Applied Probability, 416 p. Chapman & Hall Ltd., London

Aitchison J (1997) The one-hour course in compositional data analysis or compositional data analysis is simple. In: Pawlowsky-Glahn V (ed.) Proceedings of IAMG'97 – The Third Annual Conference of the International Association for Mathematical Geology, vols. I, II and addendum, pp. 3–35. International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E)

Aitchison J (2001) Simplicial inference. In: Viana M, Richards D (eds) Algebraic Methods in Statistics, Contemporary Mathematics Series. American Mathematical Society, New York, NY (in press)

Ash RB (1972) Real Analysis and Probability, 476 p. Academic Press, Inc., New York, NY

Berberian SK (1961) Introduction to Hilbert Space, 206 p. Oxford University Press, New York, NY

Buccianti A, Pawlowsky-Glahn V, Barceló-Vidal C, Jarauta-Bragulat E (1999) Visualization and modeling of natural trends in ternary diagrams: a geochemical case study. In: Lippard SJ, Næss A, Sinding-Larsen R (eds) Proceedings of IAMG '99 – The Fifth Annual Conference of the International Association for Mathematical Geology, vols. I and II, pp. 139–144. Tapir, Trondheim (N)

Cuadras CM, Fortiana J, Oliva F (1997) The proximity of an individual to a population with applications in discriminant analysis. J. Classification 14: 117–136

Eynatten Hv, Pawlowsky-Glahn V, Egozcue JJ (2001) Understanding perturbation on the simplex: a simple method to better visualise and interpret compositional data in ternary diagrams. Accepted for publication in Mathematical Geology

Fréchet M (1948) Les éléments aléatoires de nature quelconque dans une espace distancié. Annales de L'Institut Henri Poincaré 10(4): 215–308

Martín-Fernández JA, Bren M, Barceló-Vidal C, Pawlowsky-Glahn V (1999) A measure of difference for compositional data based on measures of divergence. In: Lippard SJ, Nss A, Sinding-Larsen R (eds) Proceedings of IAMG '99 – The Fifth Annual Conference of the International Association for Mathematical Geology, vols. I and II, pp. 211–216. Tapir, Trondheim (N)

Pawlowsky-Glahn V, Egozcue JJ (2001) About BLU estimators and compositional data. Accepted for publication in Mathematical Geology

Rao CR (1982) Diversity: its measurement, decomposition, apportionment and analysis. Sankhya, Indian J. Stat., Series A 44: 1–22