Moyennes et fonctions de covariance pour les données compositionnelles: une approche axiomatique

Denis Allard[1] & Thierry Marchant[2]

[1] *BioSP, INRA PACA, 84914 Avignon Cedex, France; denis.allard@inra.fr*
[2] *Ghent University, Ghent, Belgique; Thierry.Marchant@UGent.be*

Résumé. Ce travail s'intéresse à la caractérisation de la tendance centrale de données compositionnelles. Par une approche axiomatique, nous établissons de nouveaux résultats sur les propriétés théoriques de la moyenne et de la fonction de covariance pour ce type de données. Nous montrons tout d'abord que la moyenne arithmétique pondérée est la seule caractéristique de tendance centrale vérifiant les axiomes de réflexivité et de stabilité marginale. Par ailleurs, les poids intervenant dans la combinaison linéaire doivent être identiques pour toutes les composantes du vecteur compositionnel. Ce résultat a des conséquences profondes sur la structure du modèle de covariance spatiale multivariée de ces données. Dans un cadre géostatistique, nous montrons alors que le modèle de covariance proportionnelle (i.e. le produit d'une matrice de covariance et d'une fonction de corrélation) est le seul modèle de covariance pour lequel les poids du krigeage de la moyenne sont identiques pour toutes les composantes. La combinaison de ces deux résultats est que, dans le cadre des statistiques spatiales, le modèle de covariance proportionnelle est le seul modèle de covariance spatiale multivariée compatible avec les axiomes de réflexivité et de stabilité marginale.

Mots-clés. Géométrie de Aitchison; tendance centrale; équations fonctionnelles; géostatistique; fonction de covariance multivariée

Abstract. This work focuses on the characterization of the central tendency of a sample of compositional data. It provides new results about theoretical properties of means and covariance functions for compositional data, with an axiomatic perspective. As a first result, it is shown that the weighted arithmetic mean is the only central tendency characteristic verifying a small set of axioms, namely reflexivity and marginal stability. Moreover, the weights must be identical for all components of the compositional vector. This result has deep consequences on the spatial multivariate covariance modeling of compositional data. In a geostatistical setting, it is shown as a second result that the proportional model of covariance functions (i.e. the product of a covariance matrix and a single correlation function) is the only model that provides identical kriging weights for all components of the compositional data. As a consequence of these two results, the proportional model of covariance function is the only covariance model compatible with reflexivity and marginal stability.

Keywords. Aitchison geometry; central tendency; functional equation; geostatistics; multivariate covariance function

1 Introduction

Il existe de nombreuses façons de définir la "moyenne" pour les donnés compositionnelles et chacune de ces définitions vérifie des propriétés différentes. Nous nous inspirons de la définition axiomatique de Kolmogorov (1930) qui définit une "moyenne" comme une application agissant sur les données $x_1, \ldots, x_n$ qui vérifie les 4 axiomes suivants: (i) l'application est continue et strictement monotone en chacune de ses variables; (ii) elle est invariante par permutation des variables; (iii) elle est réflexive; (iv) elle est associative, dans le sens où un sous-ensemble des variables peut être remplacé par sa moyenne sans changer le résultat final. Sous ces conditions, Kolmogorov a montré qu'une "moyenne" M est nécessairement de la forme

$$M(x_1, \ldots, x_n) = \phi^{-1}\left(\frac{\phi(x_1) + \cdots + \phi(x_n)}{n}\right), \tag{1}$$

où ϕ est une fonction strictement croissante, appelée fonction génératrice. Ainsi, lorsque $\phi(x) = x,\ \phi(x) = x^{-1},\ \phi(x) = \ln(x)$ M est respectivement la moyenne arithmétique, harmonique, géométrique.

Les données compositionnelles appartiennent au simplexe de dimension $p-1$:

$$\mathbb{S}^{p-1} = \{(x^1, \ldots, x^p)\} :\ x^k \geq 0 \text{ for } k = 1, \ldots, p, \text{ and } x^1 + \cdots + x^p = 1\}. \tag{2}$$

L'analyse statistique des données compositionnelles a fait l'objet de nombreux travaux dans les dernières décades. Ces données sont le plus souvent transformées en un vecteur de $p-1$ composantes de log-ratios (Aitchison, 1986). Plusieurs transformations sont possibles comme les transformations alr, clr ou ilr (Pawlowsky-Glahn et Buccianti, 2011). Ces transformations sont des bijections vers l'espace réel de dimension $p-1$, permettant ainsi d'utiliser la panoplie habituelle des méthodes statistiques. Billheimer (2001) a montré que le simplexe $\mathbb{S}^{p-1}$ équipé du produit scalaire

$$\langle \mathbf{x}, \mathbf{y} \rangle_A = \frac{1}{p} \sum_{i=1}^{p} \sum_{j=i+1}^{p} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} = \sum_{i=1}^{p} \ln \frac{x_i}{g(\mathbf{x})} \ln \frac{y_i}{g(\mathbf{y})}, \quad \mathbf{x}, \mathbf{y} \in \mathbb{S}^{p-1}, \tag{3}$$

où $g(\mathbf{x}) = [x^1 x^2 \ldots x^p]^{1/p}$ est la moyenne géométrique des composantes de $\mathbf{x}$, induit une distance, et donc une géométrie sur le simplexe, appelée géométrie de Aitchison. Toutefois, lors de la transformation inverse vers le simplexe, des comportements inattendus peuvent se produire. En effet, sur le simplexe, lorsqu'une composante augmente, une ou plusieurs autres composantes doivent diminuer de façon à maintenir la somme égale à 1. Cette contrainte a des conséquences importantes sur les propriétés de $\mathbf{M}$. Soit un ensemble de données compositionnelles $\mathbf{x}_1, \ldots, \mathbf{x}_n$ avec q variables (i.e., $\mathbf{x}_i = (x_i^1, \ldots, x_i^q)$) que l'on peut grouper en $p < q$ variables, de deux façons différentes, créant ainsi deux nouveaux jeux de données $\mathbf{y}_1, \ldots, \mathbf{y}_n$ et $\mathbf{z}_1, \ldots, \mathbf{z}_n$. On suppose en outre que la première variable est

identique dans les deux groupes, i.e., $y_i^1 = z_i^1$, $i = 1, \ldots, n$. On ne souhaite évidemment pas que la moyenne de la première variable dépende de la façon dont sont groupées les autres variables. C'est bien le cas de la moyenne arithmétique, dont on dit qu'elle vérifie une condition de stabilité marginale. Ce n'est pas le cas de la moyenne géométrique, qui ne vérifie donc pas cette propriété. Il en va de même de toutes les autres transformations utilisant les log-ratios (alr, clr, ilr).

Dans de nombreuses applications, il ne suffit pas d'avoir une estimation sans biais des différentes composantes dans l'espace transformé si la transformation inverse introduit un biais. Souvent, disposer d'une estimation sans biais de la moyenne de la proportion brute d'une composante est une nécessité absolue. Dans ce travail nous apportons des réponses à des questions comme : pour quelles "moyennes" pouvons nous garantir la stabilité marginale ? En contexte géostatistique, quelles sont les conditions pour disposer d'une estimation sans biais de la moyenne. Quelle condition faut-il imposer à la fonction de covariance ? Ce travail apporte des réponses à ces questions. Nous traitons d'abord la caractérisation de la moyenne, puis nous établissons la caractérisation de la fonction de covariance.

2 Caractérisation axiomatique de la moyenne pour les donnés compositionnelles

Considérons un échantillon de taille fixée, n, de données compositionnelles, appartenant au simplexe $\mathbb{S}^{p-1}$. En contexte spatial, cet échantillon provient d'une variable régionalisée $\mathbf{x}(\cdot)$ définie sur $\mathbb{R}^d$, échantillonnée aux sites $\mathbf{s}_1, \ldots, \mathbf{s}_n$. Nous écrirons par la suite $\mathbf{x}_i = \mathbf{x}(\mathbf{s}_i)$. On cherche à caractériser une application $\mathbf{M} = (M^1, \ldots, M^p)$, qui associe à chaque échantillon $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ un vecteur de dimension p, $\mathbf{M}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ ayant pour composantes $M^k(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, avec $k = 1, \ldots, p$. Notons pour l'instant que $M^k(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ peut dépendre de $\mathbf{x}_1, \ldots, \mathbf{x}_n$, et pas uniquement de $x_1^k, \ldots, x_n^k$. Nous ne considérons que les applications $\mathbf{M}$ telles que $\sum_{k=1}^p M^k(\mathbf{x}_1, \ldots, \mathbf{x}_n) = 1$. En géostatistique, le krigeage peut aboutir à des pondérateurs négatifs, qui à leur tour peuvent entraîner des prédicteurs en-dehors de l'intervalle $[0, 1]$ dans certaines circonstances. Pour tenir compte de ce fait, à ce stade nous n'imposons pas de restriction à $M^k(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. L'ensemble des vecteurs $\mathbf{x}$ de dimension p vérifiant $\sum_{k=1}^p x^k = 1$ sera noté $\mathcal{S}^{p-1}$. C'est un sur-ensemble du simplexe $\mathbb{S}^{p-1}$.

Dans le cadre multivarié, les axiomes de caractérisation diffèrent des axiomes de Kolmogorov. Aux axiomes de continuité et la réflexivité, nous ajoutons l'axiome de "statibilité marginale". Nous posons ainsi les trois axiomes suivants :

[C1] Reflexivité: Pour tout $\mathbf{x} \in \mathbb{S}^{p-1}$, $\mathbf{M}(\mathbf{x}, \ldots, \mathbf{x}) = \mathbf{x}$.

[C2] Stabilité marginale : pour tout $k = 1, \ldots, p$, tout $i = 1, \ldots, n$ et tout $\mathbf{x}_1, \ldots, \mathbf{x}_n$, $\mathbf{x}'_i \in \mathbb{S}^{p-1}$: si $x_i^k = x_i'^k$, alors $M^k(\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_n) = M^k(\mathbf{x}_1, \ldots, \mathbf{x}'_i, \ldots, \mathbf{x}_n)$.

[C3] Continuité: $\mathbf{M}$ est continue pour chacun des arguments.

En présence de dépendance spatiale, la symétrie n'est en général pas une propriété souhaitable, même si elle reste une option. Afin d'être complet, nous rappelons malgré tout cette condition.

[C4] Symétrie: $M^k(\mathbf{x}_1, \ldots, \mathbf{x}_n) = M^k(\mathbf{x}_{\sigma(1)}, \ldots, \mathbf{x}_{\sigma(n)})$ pour toute permutation σ de $\{1, \ldots, n\}$ et pour tout $\mathbf{x}_1, \ldots, \mathbf{x}_n$ de $\mathbb{S}^{p-1}$.

Nous pouvons maintenant énoncer notre premier résultat. En partie A nous n'imposons pas que les composantes de $\mathbf{M}(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ soient positives (i.e. $\mathbf{M} : (\mathbb{S}^{p-1})^n \to \mathcal{S}^{p-1}$). Dans la partie B, la contrainte de positivité est imposée (i.e. $\mathbf{M} : (\mathbb{S}^{p-1})^n \to \mathbb{S}^{p-1}$).

Théorème 1 (Caractérisation de la moyenne) *Soit un échantillon de données compositionnelles $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ avec $p \geq 3$ et $n \geq 2$.*

A) L'application $\mathbf{M} : (\mathbb{S}^{p-1})^n \to \mathcal{S}^{p-1}$ vérifie les conditions C1-C3 si et seulement si, $\forall k \in \{1, \ldots, p\}$, $M^k(\mathbf{x}_1, \ldots, \mathbf{x}_{n,}) = \sum_{i=1}^n \lambda_i x_i^k$ pour des nombres réels $(\lambda_1, \ldots, \lambda_n)$ vérifiant $\sum_{i=1}^n \lambda_i = 1$.

B) L'application $\mathbf{M} : (\mathbb{S}^{p-1})^n \to \mathbb{S}^{p-1}$ vérifie les conditions C1-C2 si et seulement si, $\forall k \in \{1, \ldots, p\}$, $M^k(\mathbf{x}_1, \ldots, \mathbf{x}_{n,}) = \sum_{i=1}^n \lambda_i x_i^k$, pour des nombre réels positifs $(\lambda_1, \ldots, \lambda_n)$ vérifiant $\sum_{i=1}^n \lambda_i = 1$.

Si on impose la symétrie, alors $\lambda_i = 1/n$ pour tout $i \in \{1, \ldots, n\}$ et $\mathbf{M}(\mathbf{x}_1, \ldots, \mathbf{x}_n) \in \mathbb{S}^{p-1}$.

Les ingrédients clés de la démonstration sont la stabilité marginale et la somme à 1 (Allard et Marchant, 2018). La stabilité marginale impose que pour tout $k = 1, \ldots, p$, M^k ne dépende que de $x_1^k, \ldots, x_n^k$. Dès lors, la somme à 1 implique que la fonction génératrice de Kolmogorov est la fonction indentité, et donc que les seules applications admissibles sont des combinaisons linéaires. Il faut noter que si l'application est bornée (partie B) la continuité (Axiome C3) n'est pas nécessaire. Le théorème 1 impose qu'il y ait au moins trois composantes. Si il n'y a que deux composantes, il n'existe qu'une variable indépendante, et de nombreuses autres "moyennes" sont admissibles.

3 Données compositionnelles géostatistiques

La condition $\sum_{i=1}^n \lambda_i = 1$ correspond à une condition de non-biais. Le théorème 1 énonce donc que les moyennes des données compositionnelles satisfaisant les axiomes C1-C3 correspondent à des estimateurs linéaires sans biais. Cependant, ce théorème ne donne

aucun critère pour choisir les pondérateurs optimaux. En particulier, il n'est fait aucune référence explicite aux coordonnées des données.

Sous la condition que la fonction de covariance est connue, l'estimateur linéaire optimal sans biais est l'estimateur du moindre carré généralisé, également connu sous le nom de "krigeage de la moyenne" (Cressie, 1993). Selon le Théorème 1, les pondérateurs doivent être égaux pour toutes les composantes, ce qui implique des contraintes sur le modèle de fonction de covariance multivarié. On fait l'hypothèse que les données $(\mathbf{x}_1, \ldots, \mathbf{x}_{n,})$ proviennent d'un champ aléatoire multivarié $\mathcal{X}(\mathbf{s})$ avec $\mathbf{s} \in \mathbb{R}^d$, dont la matrice de fonctions de covariance multivariée s'écrit

$$\begin{pmatrix} C_{11}(\mathbf{h}) & \cdots & C_{1p}(\mathbf{h}) \\ \vdots & \ddots & \vdots \\ C_{p1}(\mathbf{h}) & \cdots & C_{pp}(\mathbf{h}) \end{pmatrix}, \ \mathbf{h} \ \in \mathbb{R}^d, \quad \text{avec } p \geq 3, \text{et } n \geq 2. \tag{4}$$

Pour un ensemble de n sites $(\mathbf{s}_1, \ldots, \mathbf{s}_n)$, ce modèle implique une matrice-bloc $np \times np$ de la forme

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{11} & \cdots & \mathbf{C}_{1p} \\ \vdots & \ddots & \vdots \\ \mathbf{C}_{p1} & \cdots & \mathbf{C}_{pp} \end{pmatrix}, \tag{5}$$

où chaque matrice $\mathbf{C}_{kl}$ est telle que $[\mathbf{C}_{kl}]_{ij} = C_{kl}(\mathbf{s}_j - \mathbf{s}_i)$, avec $1 \leq k, l \leq p$ et $1 \leq i, j \leq n$.

Théorème 2 *Soit un champ aléatoire multivarié ($p \geq 2$) stationnaire d'ordre 2.*

A) $\forall n \geq 2$ *et pour tout* $(\mathbf{s}_1, \ldots, \mathbf{s}_n)$*, les pondérateurs des p moyennes sont égaux si et seulement si la fonction de covariance multivariée suit un modèle proportionnel, i.e., pour* $1 \leq k, l \leq p$*,*

$$C_{kl}(\mathbf{h}) = \sigma_{kl}\rho(\mathbf{h}), \ \mathbf{h} \in \mathbb{R}^d,$$

pour une matrice de covariance $\mathbf{\Sigma} = [\sigma_{kl}]_1^p$ *et une fonction de corrélation* $\rho(\mathbf{h})$*.*

B) Ce résultat reste valable si on impose aux pondérateurs d'être positifs.

Ce théorème est démontré dans Allard et Marchant (2018). Considéré avec le théorème 1, cela implique que pour l'estimation d'une "moyenne" de données compositionnelles, le seul estimateur de cette moyenne sans biais, de variance minimale et vérifiant la réflexivité et la stabilité marginale est le krigeage de la moyenne (y compris lorsqu'on impose des contraintes de positivité sur les pondérateurs) utilisant un modèle de covariance multivariée qui suit un modèle proportionnel $\mathbf{C}(\mathbf{h}) = \mathbf{\Sigma} \otimes \rho(\mathbf{h})$.

4 Discussion

Ces résultats montrent que dans l'analyse de données compositionnelles (en contexte spatialisé ou non), il n'est pas possible, en même temps, de vérifier les axiomes C1-C3 et d'utiliser la modélisation à l'aide des transformations log-ratio, ni des fonctions de covariances multivariées en dehors du modèle proportionnel. C'est un résultat d'impossibilité, forcément frustrant, mais dont il faut s'accommoder. D'un côté, imposer la stabilité marginale alors qu'un modèle plus complexe que celui du théorème 2 est vrai peut mener à des estimateurs peu efficaces. De l'autre, utiliser un modèle complexe peut mener à la perte de la stabilité marginale, pouvant induire des comportements parfois contre-intuitifs comme illustrés dans Allard et Marchant (2018). Selon le cas d'étude et la qualité des données, il sera préférable de renoncer à la condition de non-biais ou de renoncer aux transformations log-ratio.

Bibliographie

Aitchison, J. (1986) *The statistical analysis of compositional data*. Chapman and Hall

Allard, D. and Marchant, T. (2018) Means and covariance functions for geostatistical compositional data: an axiomatic approach. *Mathematical Geosciences*, 50: 299-315.

Billheimer, D., Guttorp, P. and Fagan, W.F. (2001) Statistical Interpretation of Species Composition. *Journal of the American Statistical Association* 96:1205–1214

Cressie, N. (1993) *Statistics for Spatial Data, Revised Edition*. Wiley

Kolmogorov, A. (1930) Sur la notion de la moyenne. *Atti Accad. Naz. Lincei* . 12, 88?391

Pawlowksy-Glahn, V. and Olea, R.A. (2004) *Geostatistical Analysis of Compositional Data.* Oxford University Press

Pawlowsky-Glahn, V. and Buccianti, A. (2011). *Compositional Data Analysis: Theory and Applications.* Wiley,