REVUE DE LA LITTERATURE SUR L'USAGE ET MESUSAGE DES METHODES STATISTIQUES POUR L'ANALYSE DE DONNEES COMPOSITIONNELLES D'ENTRAINEMENT EN SCIENCES DU SPORT

Pauline Desnavailles[1] & Carolyn Ingram[1] & Thomas Prince[1] & Marta Avalos-Fernandez[1,2,3]

[1]*Univ. Bordeaux, ISPED, F-33000 Bordeaux, France, prenom.nom@etu.u-bordeaux.fr*
[2]*Univ. Bordeaux, Inserm, Bordeaux Population Health Research Center, UMR1219, F-33000 Bordeaux, France, prenom.nom@u-bordeaux.fr*
[3]*Inria SISTM Team, F-33405 Talence, France, prenom.nom@inria.fr*

Résumé. Les données de composition ou données compositionnelles (CoDa pour *Compositional Data*) consistent en des parts : leur somme vaut 100% pour chaque sujet, elles véhiculent une information relative. Puisqu'une composante peut être déterminée à partir de la somme du reste de la composition, les composantes sont mathématiquement et statistiquement dépendantes. Cette structure complique l'analyse et ne permet pas d'effectuer des inférences valides à partir d'analyses statistiques classiques. Aitchison, 1982 et Egozcue et collègues, 2003, entre autres, ont fourni un cadre pour analyser des CoDa en projetant les données de l'espace simplex contraint à l'espace euclidien en utilisant des transformées non linéaires telles que la log-cote. Cependant, même après transformation, la nature compositionnelle reste inhérente aux données. Le traitement statistique des CoDa est ainsi loin d'être un exercice aisé. Dans certains domaines d'application, les CoDa sont couramment utilisées alors que leur nature est ignorée. De nombreux tutoriels ont ainsi été proposés afin de guider l'analyse tout en tenant compte des spécificités du champ d'application. Qu'en est-il en sciences du sport ? L'objectif de ce travail est de faire un état des lieux sur le traitement statistique des CoDa dans un cadre précis, celui de l'analyse de la distribution de l'entraînement en termes d'intensité.

Mots-clés. Statistique et Sport, Données compositionnelles, Analyse de l'entraînement, Revue de la littérature

Abstract. Compositional data (CoDa) consist of parts: their sum is 100% for each subject, they carry relative information. Since a component can be determined from the sum of the rest of the composition, the components are mathematically and statistically dependent. This structure complicates the analysis and valid inferences cannot be drawn from standard statistical analyses. Aitchison, 1982 and Egozcue and colleagues, 2003, among others, provided a framework for analyzing CoDa by projecting data from the simplex constrained space to the Euclidean space using non-linear transformations such as log-ratio. However, even after transformation, the compositional nature remains inherent in the data. The statistical treatment of CoDa is thus far from being an easy exercise. In some applied sciences, CoDa are commonly used while its nature is ignored. Many tutorials were thus proposed to guide the analysis while taking into account the specificities of the field. What about sports sciences? The objective of this work is review the statistical treatment of CoDa in a precise framework, that of the analysis of the training intensity distribution.

Keywords. Statistics and Sport, Compositional data, Training analysis, Literature review

1. Introduction

Une des dimensions de la discipline de la statistique est celle d'outil de description, d'analyse et de décision que l'ensemble des disciplines scientifiques basées sur la recherche quantitative se sont approprié suivant leurs spécificités (Jutand, 2015). Le sport, quant à lui, est une discipline qui se prête naturellement à la quantification (Renaud, 2015 ; Foulley, 2017). Un exemple est donné par l'analyse quantitative de l'entraînement et la performance sportive en vue de l'amélioration de cette dernière.

Atteindre une performance sportive optimale à des moments précis (à savoir, les compétitions majeures) est le principal objectif des programmes d'entraînement des sportifs professionnels. L'amélioration de la performance, tout en limitant les risques de surentraînement et de blessure, exige une gestion minutieuse des paramètres de l'entraînement : volume, fréquence, type, intensité… Le volume d'entraînement, mesuré en termes de distance ou de durée, est le premier paramètre monitoré par les entraîneurs. L'intensité d'entraînement ou charge de travail permet aux entraîneurs de monitorer le degré d'effort. La mesure de ce paramètre, particulièrement important dans les sports d'endurance, est moins directe. Souvent des quantifications sont effectuées à l'aide de la fréquence cardiaque, du taux de lactate dans le sang ou du débit maximum d'oxygène consommé qui permettent, ensuite, de classer l'exercice physique selon trois zones d'intensité : basse, moyenne, ou haute intensité (ou, plus finement, selon de multiples zones définies en relation avec les notions de capacité aérobie et anaérobie et de travail au seuil d'accumulation d'acide lactique) (Laursen et Buchheit, 2019).

Des travaux en sciences du sport visent l'élaboration de recommandations pour le monitoring de l'entraînement à partir de l'étude de la distribution de l'intensité (Sylta et *al.*, 2016). Ces études évaluent des questions telles que « Est-ce que les distributions de l'intensité observées dans ces différents groupes de sportifs ou ces différentes conditions expérimentales sont comparables ? » (Figure 1, à gauche) ou « Quelle distribution de l'intensité, parmi celles observées, contribue à une meilleure performance ? ». Analyser la distribution de l'entraînement en termes d'intensité, abstraction faite du volume (habituellement personnalisé sur-mesure), revient à normaliser les données de chaque sujet en les divisant par son volume individuel total. On calcule ainsi des proportions ou des pourcentages des intensités d'entraînement.

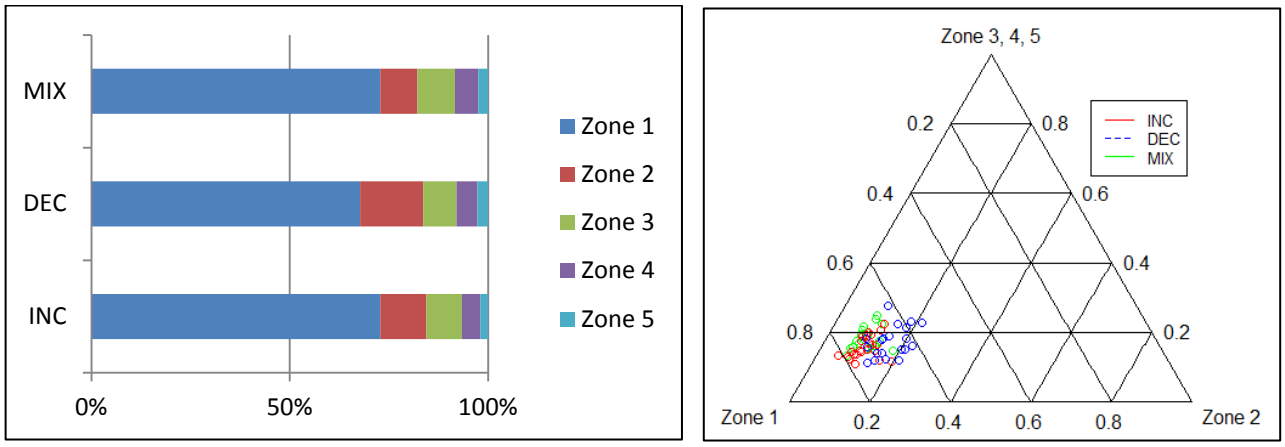

Figure 1. A gauche : Distribution de l'intensité moyenne dans trois groupes selon un entraînement à haute intensité croissant (INC), décroissant (DEC) ou mixte (MIX) et selon un classement de l'intensité en 5 zones (1, 2, 3, 4 et 5 définies par un effort de 60%–75%, 75%–85%, 85%–90%, 90%–95% et 95%–100% de la fréquence cardiaque maximale, respectivement). Données extraites de Sylta et *al.*, 2016. A droite : diagramme ternaire en regroupant les zones 3, 4 et 5. Données simulées à partir d'une distribution Dirichlet sur le simplex, les valeurs des paramètres sont extraits de Sylta et *al.*, 2016.

En tant que pourcentages, les intensités d'entraînement ne peuvent pas être considérées isolément. Un pourcentage élevé d'une intensité d'entraînement comporte automatiquement des pourcentages plus faibles d'autres intensités. Puisqu'une composante peut être déterminée à partir de la somme du reste de la composition, les composantes sont, en effet, mathématiquement et statistiquement dépendantes les unes des autres. Il s'agit de données compositionnelles ou données de composition (CoDa pour *Compositional Data*) (Aitchison, 1982, 1986 ; Greenacre, 2018). Les CoDa consistent en des parts : leur somme vaut 100% pour chaque sujet, elles véhiculent une information relative. Formellement, un vecteur de dimension *D*, $\boldsymbol{x} = (x_1, \dots, x_D)^t$, est compositionnel lorsque les *D* compositions ou parts vérifient : $x_i > 0,\ i = 1, \dots, D,$ et $\sum_{i=1}^{D} x_i = 100\%$. L'ensemble de ces vecteurs constituent le simplex $\mathbb{S}^D$, sous-espace de dimension *D-1* de $\mathbb{R}^D$. La figure 1 (à droite) montre la représentation de données dans $\mathbb{S}^3$.

La structure compositionnelle induit des corrélations entre les composantes, complique l'analyse car les analyses statistiques classiques ne permettent pas d'effectuer des inférences valides, et rend délicate l'interprétation des résultats.

Par exemple, dans le cas de la corrélation, cette mesure est basée sur des variances et des covariances qui sont définies pour l'espace euclidien et non pour le simplex. En effet, la covariance nous amène au produit scalaire euclidien des variables centrées et, la variance, à la distance quadratique à la moyenne. De nombreuses techniques d'analyse de données multivariées, telle que l'analyse en composantes principales, reposent sur ces mesures. De plus, sur le simplex, un biais négatif est présent dans la structure de covariance, représenté par la relation : $\mathrm{cov}(x_1, x_2) + \mathrm{cov}(x_1, x_2) + \cdots + \mathrm{cov}(x_1, x_D) = -\mathrm{var}(x_1)$. Des propriétés non souhaitables en découlent, comme la dépendance de l'échelle ou l'incohérence sous-compositionnelle (à savoir, si on s'intéresse à un sous-ensemble des composantes, normalisé à nouveau pour que la somme soit 100%, il n'y aura aucune relation entre la structure de covariance de cette sous-composition et celle de la composition complète).

En ce qui concerne les tests de comparaison de la distribution entre deux groupes, si pour un groupe nous observons des pourcentages plus élevés d'une intensité d'entraînement, nous observerons forcément des pourcentages plus élevés d'au moins une intensité différente pour l'autre groupe. Tester la différence de la distribution (globalement), sans tenir compte de l'interdépendance des pourcentages, peut conduire à des résultats faussement significatifs. Une alternative consiste à tester une seule composante à la fois. Pourtant, l'analyse d'une seule composante dans un ensemble de données composites ne répond pas, en général, à la question posée et ne permet pas de comparer dans son ensemble la distribution des intensités d'entraînement entre les groupes. Une autre alternative consisterait à se tourner vers l'analyse de variance multivariée (ou MANOVA pour *Multivariate analysis of variance*). Toutefois, cette technique suppose que les données sont normalement distribuées et une telle hypothèse est invalide pour les données compositionnelles (Aitchison, 1986).

Aitchison (1982, 1986) et Egozcue et *al.* (2003), entre autres, ont fourni un cadre pour analyser des CoDa en projetant les données de l'espace simplex contraint à l'espace euclidien en utilisant des transformées non linéaires telles que la log-cote (additive, centrée ou isométrique, parmi les plus utilisées). Souvent, une transformation logarithmique permet de rendre la distribution symétrique, voire normale. Cependant, les résultats demeurent dépendants des métriques utilisées (Avalos et *al.*, 2018), et même après transformation, la nature compositionnelle reste inhérente aux données.

Le traitement statistique des CoDa est ainsi loin d'être un exercice aisé. Des tutoriels ont été proposés afin de guider l'analyse statistique des CoDa en tenant compte des spécificités du champ d'application (Campbell et *al.*, 2009 ; Pennington et *al.*, 2009 ; Mert et *al.*, 2018). Qu'en est-il en sciences du sport ? L'objectif de ce travail est de faire un état des lieux sur le traitement statistique des CoDa. Pour cela, nous avons effectué une revue de la littérature des travaux portant sur des analyses de la distribution de l'entraînement en termes d'intensité.

2. Méthodologie

Dans cette section, nous détaillons et justifions les étapes de notre revue de la littérature. Dans les sciences appliquées, la nature compositionnelle des données est souvent ignorée. Il n'est donc pas judicieux de fonder notre recherche bibliographique sur les mots clés « *compositional data* » et « *sport(s)* ». En effet, un seul article dans le domaine des sciences du sport pointe la nature compositionnelle des données (Ordóñez *et al*., 2016).

Les moteurs de recherche Pubmed-Medline (base de données bibliographiques en sciences de la vie, médecine et sciences biomédicales), SPORTDiscus (base de données bibliographiques, texte intégral compris, en sport et éducation physique) et Scopus (base de données bibliographiques multidisciplinaire) ont été utilisés.

Nous avons effectué des explorations préalables dans les différents moteurs de recherche ainsi que dans le thésaurus MeSH (pour Medical Subject Headings, thésaurus de référence dans le domaine biomédical). Les mots-clés retenus sont les suivants :

- Mots clés concernant la distribution de l'entraînement : "training intensity distribution", "polarized training", "high-intensity interval training", "interval training" et "periodization"
- Mots clés concernant la population : "elite", "professional athletes", "competitive", "high level".

Afin de limiter notre recherche, nous avons sélectionné les 40 revues en Sciences du Sport les mieux classées en 2017 selon l'indicateur SJR (SCImago Journal Rank) de l'influence scientifique de revues académiques. La grande majorité est classée dans le premier quartile du rang. Nous avons exclu les revues dont le titre était focalisé sur des thématiques non pertinentes (psychologie du sport, par exemple). D'autre part, nous avons rajouté une revue non classée dans la catégorie de Sciences du Sport, mais souvent citée dans la bibliographie des articles trouvés.

Les revues sélectionnées sont : *American Journal of Sports Medicine, Sports Medicine, British Journal of Sports Medicine, Medicine and Science in Sports and Exercise, Journal of Physiology, International Journal of Sports Physiology and Performance, Journal of Science and Medicine in Sport, Scandinavian Journal of Medicine and Science in Sports, Journal of Applied Physiology, Journal of Athletic Training, Journal of Strength and Conditioning Research, International Journal of Sports Medicine, European Journal of Applied Physiology, Journal of Sports Sciences, European Journal of Sport Science, Applied Physiology, Nutrition and Metabolism, Journal of Sports Science and Medicine, Frontiers in Physiology*.

Ainsi, afin d'obtenir les articles correspondant aux mots-clés établis et publiés dans les journaux sélectionnés, nous avons utilisé (dans l'ensemble des bases) l'algorithme suivant:

(("training intensity distribution" OR "polarized training" OR "high-intensity interval training" OR "interval training" OR periodization) AND (elite OR "professional athletes" OR competitive OR "high level")) AND (("Journal Of Strength And Conditioning Research") OR "International Journal Of Sports Physiology And Performance") OR ("International Journal Of Sports Medicine") OR ("Frontiers In Physiology") OR ("Medicine And Science In Sports And Exercise") OR ("Journal Of Sports Science And Medicine") OR ("Journal Of Sports Sciences") OR ("European Journal Of Applied Physiology") OR ("European Journal Of Sport Science") OR ("Scandinavian Journal Of Medicine And Science In Sports") OR ("Journal Of Science And Medicine In Sport") OR ("American Journal Of Sports Medicine") OR ("Journal Of Applied Physiology") OR ("Applied Physiology Nutrition And Metabolism") OR ("British Journal Of Sports Medicine") OR ("Journal Of Physiology"))

Sur base du titre ou du résumé, dans un premier temps, et sur base du texte complet, dans un deuxième temps, nous avons exclu les articles ne portant pas sur la distribution de l'intensité d'entraînement, les articles ne concernant pas les sportifs professionnels, les sujets médicaux et les revues de la littérature.

3. Résultats

Au total, 515 références ont été récupérées grâce à nos équations de recherche, 330 après élimination des doublons. Parmi eux, 205 articles ne portant pas sur la distribution de l'intensité de l'entraînement et puis, 12 articles ne concernant pas les sportifs professionnels ont été éliminés. Nous avons ensuite exclu 70 articles dont le sujet principal portait sur l'aspect médical de l'entraînement. Enfin, nous avons exclu 7 revues de la littérature. La lecture du texte intégral de 35 références a été effectuée, ce qui a conduit à l'exclusion de 15 références supplémentaires dont le contenu ne correspondait finalement pas au sujet défini par nos critères. La figure 2 montre le diagramme de flux.

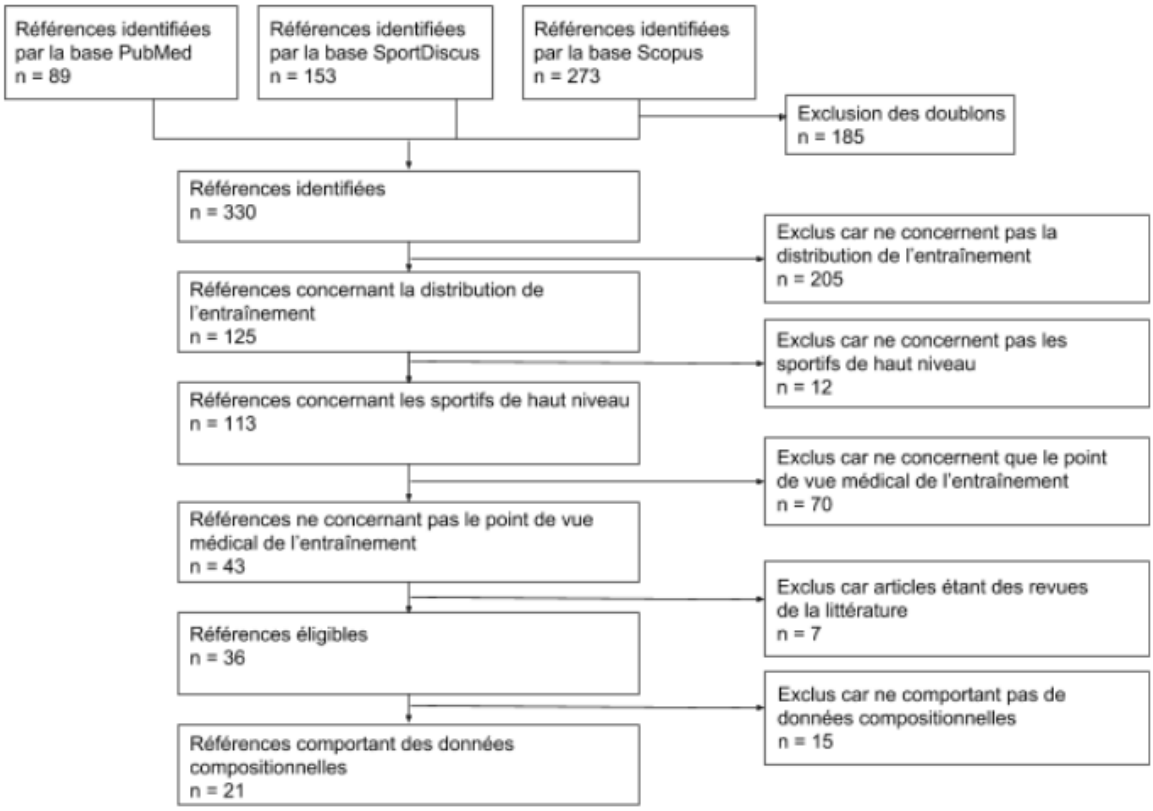

Figure 2 : Diagramme de flux illustrant le processus de sélection des études de la revue bibliographique. Date : 10/02/2019.

4. Conclusion

Dans cette présentation, nous analyserons les traitements statistiques qui ont été appliqués aux données compositionnelles dans les études sélectionnées lors de notre revue de la littérature, ainsi que les implications sur les conclusions avancées, lorsque ces traitements s'avèrent inappropriés.

Remerciements

Ce travail a été effectué dans le cadre d'un projet tutoré de la 1ère année du Master Santé Publique, modalité en présentiel, parcours Biostatistique, de l'ISPED, Université de Bordeaux. Nos remerciements vont à Valérie Kiewsky, responsable pédagogique de cette formation, pour sa relecture attentive.

Bibliographie

Aitchison, J. (1982), The Statistical Analysis of Compositional Data. *J. R. Statist. Soc. B*, 44(2), pp. 139-177.

Aitchison, J. (1986), The Statistical Analysis of Compositional Data. Chapman & Hall, Ltd. London, UK.

Avalos, M., Nock, R., Ong, C.S., Rouar, J., et Sun, K. (2018), Representation Learning of Compositional Data. In: S. Bengio; H. Wallach; H. Larochelle; K. Grauman; N. Cesa-Bianchi; R. Garnett. Thirty-second Conference on Neural Information Processing Systems, Dec 2018, Montréal, Canada. Advances in Neural Information Processing Systems 31.

Campbell, G.P., Curran, J.M., Miskelly, G.M., Coulson S., Yaxley, G.M., Grunsky, E.C., et Cox, S.C. (2009), Compositional data analysis for elemental data in forensic science, *Forensic Science International*, 188, pp.81–90.

Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G. et Barcelo-Vidal, C. (2003), Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, 35(3), pp. 279-300.

Foulley, J.-L. (2017) Pour la mathématique statistique dans le sport. Variances.eu le webzine des Ensae Alumni. Dossier Sport. Entretien avec Tassi, P. avril 2017. url : http://variances.eu/?p=2244 [Consulté le 2/2/2019]

Greenacre, M. (2018), Compositional Data Analysis in Practice. Chapman and Hall/CRC, Boca Raton, FL, USA.

Jutand, M.A. (2015), Études des phénomènes de transposition didactique de la statistique dans le champ universitaire et ses environnements : une contribution à la pédagogie universitaire. Thèse de doctorat en éducation. Université de Bordeaux.

Laursen, P. et Buchheit, M. (2018), Science and Application of High-Intensity Interval Training, Human Kinetics, Champaign, USA.

Mert, M.C., Filzmoser, P., Endel, G. et Wilbacher, I. (2018) Compositional data analysis in epidemiology, *Stat Methods Med Res.*, 27(6), pp.1878-1891.

Ordóñez, E.G., Pérez, M.D. et González, C.T. (2016), Performance Assessment in Water Polo Using Compositional Data Analysis, *J Hum Kinet.*, 54, pp.143-151.

Pennington, L., James, P., McNally, R., Pay, H., et McConachie, H. (2009), Analysis of compositional data in communication disorders research, *Journal of Communication Disorders*, 42 , pp.18–28.

Renaud, A. (2015), La tête et les jambes. Un cours de statistique au pas de course ? Colloque francophone international sur l'enseignement de la statistique (CFIES'2015), Bordeaux

Sylta, Ø., Tønnessen, E., Hammarström, D., Danielsen, J., Skovereng, K., Ravn, T., Rønnestad, B.R., Sandbakk, Ø., et Seiler, S. (2016), The effect of different high-intensity periodization models on endurance adaptations. *Med Sci Sports Exerc.*, 48(11), pp.2165-2174.