Comparaison de méthodes d'analyses multivariées pour la description de données de germination de semences

O. Thierry [a], R. Boumaza [a], J. Buitink [a], C. Landès [a], O. Leprince [a], P. Santagostini [a] and J. Bourbeillon [a]

[a] IRHS, Agrocampus-Ouest, INRA, Université d'Angers, SFR 4207 QuaSaV
42 rue Georges Morel, 49071 Beaucouzé Cedex, France
prenom.nom@inra.fr

Mots clefs : Analyse en composantes principales, STATIS dual, Densité de probabilité, Corrélation, Biologie végétale.

L'émergence de technologies d'imagerie (acquisition et analyse d'images) en biologie a permis un changement d'échelle dans le nombre d'individus pouvant être observés conjointement et dans la fréquence à laquelle les mesures peuvent être réalisées. Par exemple, dans les études de germination de semences, là où un nombre réduit de lots, contenant un nombre réduit de graines, pouvaient être évalués, à des intervalles de temps assez longs, ce sont des centaines de lots incluant des centaines de graines pour lesquels des descripteurs peuvent être extraits à un rythme qui se rapproche de plus en plus du temps réel. Pour le biologiste, ce changement d'échelle pose un véritable challenge pour capturer et décrire finement les comportements des lots de semences. Il représente également une véritable opportunité non encore exploitée de prendre en compte l'hétérogénéité des individus constituant les lots.

Préalablement à l'analyse de l'effet d'un ou plusieurs facteurs (génotype, traitement, etc.) sur les différentes variables quantitatives acquises lors des expériences, les analyses de corrélation et analyses multivariées telles que l'analyse en composantes principales (ACP) [1] sont désormais employées en routine par les biologistes. Dans le contexte d'acquisitions à grande échelle structurées en lots, une pratique usuelle consiste à représenter chaque lot par une valeur calculée unique pour chacune des variables mesurées (moyenne, médiane, etc.). Cette approche permet de simplifier l'analyse et améliorer la lisibilité de la visualisation en réduisant le nombre de points représentés. L'hétérogénéité au sein d'un lot pourrait toutefois être potentiellement caractéristique du comportement de certains systèmes biologiques, telles que les semences, et être liée à d'autres facteurs, identifiés ou non lors de la réalisation de l'analyse. Il est donc important de voir à quel point il pourrait être pertinent de prendre en compte cette hétérogénéité au niveau de l'ACP.

Afin d'évaluer les approches alternatives de représentation des lots, nous avons choisi de comparer les résultats obtenus avec 3 méthodes de type ACP sur un jeu de données réelles acquises au cours du projet ANR REGULEG. Dans le cadre de ce projet, environ 200 génotypes différents représentant la diversité génétique d'une légumineuse modèle ont été cultivés soit en conditions normales, soit en condition de stress hydrique. Les graines des plantes ont été récoltées et 100 graines par génotype par condition de culture ont été semées sur un banc de germination et photographiées toutes les 2h pendant la germination, conduisant à environ 38000 individus mesurés. Des descripteurs morphologiques pour chaque graine et pas de temps ont ensuite été extraits des photographies par analyse d'image. Une première étape de corrélation nous a permis de réduire le nombre de variables étudiées pour en conserver une dizaine décrivant l'aspect physique des graines et leurs comportements. En complément, les génotypes sont décrits par une dizaine de facteurs climatiques qui pourraient être liés aux variables mesurées sur les graines.

La comparaison porte principalement sur les sorties graphiques et les aides à l'interprétation fournies par :

- l'analyse en composantes principales [1] portant sur les moyennes des variables associées aux lots, telle qu'implémentée dans le package R FactoMineR [2];
- la méthode STATIS dual [3] sur les matrices de variance-covariance, puis sur les matrices de corrélation associées aux lots, implémentée dans le package multigroup [4] ;
- l'analyse en composantes principales de densités de probabilité associées aux lots [5], implémentée dans le package dad [6].

Remerciements
Ce travail a été partiellement financé par le projet REGULEG ANR-15-CE20-0001 et le projet RFI « Objectif Végétal » DIVIS.
Nous remercions la plateforme PHENOTIC pour l'acquisition et l'analyse des images de germination de semences.

Références

[1] Jolliffe I.T. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, 2002
[2] Husson F., Josse J., Yousfi S., Le S. and Mazet J., (2018). FactoMineR: Multivariate Exploratory Data Analysis and Data Mining. R package, version 1.40. https://cran.r-project.org/package=FactoMineR
[3] Lavit C., Escoufier Y., Sabatier R., Traissac Pierre. (1994). The ACT (STATIS method). Computational Statistics and Data Analysis, (18), 97-119.
[4] Eslami A., Qannari EM, Bougeard S. , Sanchez G., Questions, comments go to Aida Eslami and Stephanie Bougeard (2015). multigroup: Multigroup Data Analysis. R package version 0.4.4. https://CRAN.R-project.org/package=multigroup
[5] Boumaza, R., Yousfi, S., Demotes-Mainard, S. (2015). Interpreting the principal component analysis of multivariate density functions. Communications in statistics. Theory and methods, 44(16): 3321-3339.
[6] Boumaza R., santagostini P., Yousfi S., Hunault G., Bourbeillon J., Pumo B. and Demotes-Mainard S. (2018). dad: Three-Way Data Analysis Through Densities. R package, version 3.1.0. https://CRAN.R-project.org/package=dad