Analyse de $(K+1)$ tableaux avec le logiciel `ade4`. Application en épidémiologie.

S. Bougeard[a] and **S. Dray**[b]

[a]Département d'épidémiologie
Anses (Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail)
Zoopôle, 22440 Ploufragan, France
stephanie.bougeard@anses.fr

[b]Laboratoire de biométrie et biologie évolutive
CNRS - Université Lyon 1
UMR CNRS 5558, 43 bd du 11 novembre 1918, 69622 Villeurbanne, France
stephane.dray@univ-lyon1.fr

Mots clefs : Statistique, Biologie, Régression multi-tableaux, `Ade4`.

Dans de nombreux domaines, sont recueillies des données présentant une structure en plusieurs tableaux dont il convient de tenir compte à la fois pour le traitement statistique mais aussi pour l'interprétation. Nous traitons ici du cas où celles-ci sont organisées en $(K+1)$ tableaux, à savoir un tableau Y comprenant plusieurs variables à expliquer et K tableaux $(X_1, \ldots, X_K)$ comprenant chacun plusieurs variables explicatives, l'ensemble de ces variables étant mesuré sur les mêmes observations. Les variables sont supposées quantitatives, mais des variables qualitatives peuvent être intégrées après codage disjonctif. Dans la bibliographie, les principaux domaines dans lesquels ces données sont décrites sont le suivi de processus industriels, la chimiométrie, l'analyse sensorielle, les études de marché, l'écologie et l'épidémiologie (*e.g.* Figure 1).

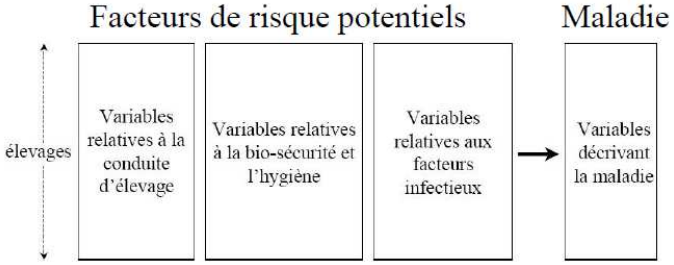

Figure 1: Exemple de données d'épidémiologie vétérinaire organisées en $(K+1)$ tableaux.

Les données multi-tableaux étant complexes, les objectifs poursuivis peuvent être multiples. Dans les domaines décrits précédemment, l'objectif majeur est de réaliser une description ainsi qu'une prédiction du tableau Y à partir des K tableaux explicatifs $(X_1, \ldots, X_K)$. Le traitement statistique de ce type de données pose actuellement problème car aucune méthode adaptée n'est actuellement implémentée dans des logiciels, qu'ils soient libres ou commerciaux. Deux types de solutions peu satisfaisantes pour l'utilisateur restent possibles : (i) simplifier la structure de ces données et appliquer une méthode développée pour deux tableaux Y et X, *e.g.* la régression PLS (package `pls` de `R`) ou l'analyse des redondances (*e.g.*, fonction `pcaiv` du package `ade4`), (ii) ou au contraire utiliser des méthodes développées pour des données plus complexes, *e.g.* l'approche PLS (package `plspm` de `R`) avec l'inconvénient d'un algorithme complexe dont la convergence n'est pas démontrée. Le logiciel d'analyse de données `ade4` [3] propose un ensemble de méthodes

d'analyse de données pour le traitement d'un seul, de deux mais aussi de K tableaux [2].

Afin de proposer aux utilisateurs des méthodes adaptées au traitement de $(K+1)$ tableaux, deux d'entre elles ont été récemment développées et implémentées dans le logiciel `ade4` : la régression PLS multibloc [4, 5] et l'analyse en composantes principales sur variables instrumentales multibloc, aussi appelée analyse des redondances multibloc [1]. La régression PLS multibloc (fonction `mbpls`) a été choisie pour sa popularité et sa stabilité pour le cas de variables explicatives nombreuses et corrélées, mais il est démontré que pour le cas d'un seul tableau Y ses principaux résultats sont ceux de la régression PLS classique. L'analyse en composantes principales sur variables instrumentales multibloc (fonction `mbpcaiv`) a été choisie pour sa bonne adaptation aux données structurées en $(K+1)$ tableaux ayant une visée prédictive, mais peut présenter des limites en cas de quasi-colinéarité marquée au sein des tableaux explicatifs. Pour pouvoir utiliser ces deux fonctions, il convient de définir le tableau Y comme un objet de la classe `dudi` (classe d'objet `ade4` pour les données organisées en un seul tableau) et le tableau X comme un objet `ktab` (classe d'objet `ade4` pour les données structurées en K tableaux). En complément et dans l'objectif de valoriser au mieux les nombreux résultats issus des méthodes multi-tableaux, des outils d'aide à l'interprétation pour la description mais aussi pour l'explication sont spécifiquement développés. (i) Du point de vue descriptif, la fonction `summary` fournit pour chaque dimension, l'inertie et la variance expliquée de chaque tableau par les variables latentes. Pour entrer dans le détail des liens entre variables et entre tableaux en lien avec les observations, des représentations factorielles graphiques sont proposées par la fonction `plot`. (ii) Du point de vue explicatif, la fonction `testdim` permet à l'utilisateur de choisir la dimension optimale du modèle par validation croisée. Une fois cette dimension définie, il est possible de calculer et de représenter les intervalles de confiance des principaux indices d'aide à l'interprétation, *i.e.* coefficients de régression, importance globale des variables explicatives et importance globale des tableaux explicatifs.

Une application est proposée en épidémiologie vétérinaire. Les données proviennent d'une enquête analytique menée sur un échantillon de 351 lots de poulets. L'objectif est de comprendre les facteurs de risques globaux des pertes (Y) décrites par quatre variables, *i.e.* la mortalité durant la première semaine, la mortalité durant le reste de la période d'élevage, la mortalité pendant le ramassage et le transport, le taux de saisie à l'abattoir. Les variables explicatives sont organisées en 4 tableaux, *i.e.*, X_1 relatif à la structure de l'élevage, X_2 aux caractéristiques du lot la première semaine, X_3 aux caractéristiques du lot durant le reste de l'élevage et X_4 au ramassage, transport et abattage. La finalité pour l'épidémiologiste est d'avoir une vision globale des actions à mener dans les différentes phases de production afin de réduire les pertes.

Références

[1] Bougeard, S., Qannari, E.M., Rose, N. (2011). Multiblock Redundancy Analysis: interpretation tools and application in epidemiology. *Journal of Chemometrics*, **25**(9), 467-475

[2] Dray, S., Dufour, A.B., Chessel., D. (2007). The `ade4` package - II: Two-table and K-table methods. *R News*, **7**(2), 47-52

[3] Dray, S., Dufour, A.B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, **22**(4):1-20.

[4] Wangen, L.E., Kowalski, B.R. (1988). A multiblock partial least squares algorithm for investigating complex chemical systems. *Journal of Chemometrics*, **3**, 3-20

[5] Wold, S. (1984). Three PLS algorithms according to SW. *Symposium MULDAST (Multivariate analysis in science and technology)*, Umea University, Sweden, 26-30