DÉVELOPPEMENT DE MÉTHODES DE RÉGRESSION MULTIBLOC POUR LE TRAITEMENT DES DONNÉES D'ÉPIDÉMIOLOGIE ANIMALE. Des besoins statistiques des épidémiologistes au développement d'un package R *

Stéphanie Bougeard[1], **Coralie Lupo**[2], **Claire Chauvin**[1], **Christelle Fablet**[1], **Stéphane Dray**[3] et **Nicolas Rose**[1]

RÉSUMÉ

L'objectif de cet article est d'illustrer la démarche de développement de méthodes statistiques en vue de leur application. Un exemple est proposé pour le développement de méthodes de régression multibloc à variables latentes appliquées aux données d'épidémiologie animale. Les différentes étapes de cette démarche sont illustrées : la mise en place du cahier des charges auquel ces méthodes doivent répondre, le développement méthodologique, le développement d'indices et de graphes d'aide à l'interprétation, la phase de test avec l'application à de nombreux jeux de données et enfin le développement d'un package sur le logiciel libre R permettant un accès aisé pour tous à ces méthodes. Une illustration à partir d'une méthode de régression multibloc, ainsi que de son mode d'utilisation par le package R développé, est proposée sur des données d'épidémiologie analytique dont l'objectif est de déterminer les facteurs de risque des pertes (mortalité et saisies) dans la filière poulet de chair.

Mots-clés : épidémiologie analytique, facteur de risque, analyse factorielle multibloc, analyse des redondances, package R.

SUMMARY

The purpose of this paper is to describe the development strategy of statistical methods to be applied in the field. As an example, the development of multiblock regression methods applied to veterinary epidemiological data is described. The sequential steps in this strategy are detailed: listing of constraints to be considered, methodological development, associated indexes and design of graphical displays, testing phase with applications to numerous datasets and finally package development on the free R software to make these methods available to all potential users. An illustration of a multiblock regression method and of way how to apply it to epidemiological data in broiler chickens using the R package is presented in order to identify risk factors of losses (mortality and condemnation).

Keywords: Epidemiology, Risk factor, Multiblock method, Redundancy analysis, R package.

* Texte de la communication orale présentée au cours des Journées scientifiques AEEMA, 31 mai 2013

[1] Université européenne de Bretagne - Anses, Laboratoire de Ploufragan-Plouzané - Unité d'épidémiologie et bien-être du porc, BP 53, Technopole Saint Brieuc Armor, 22440 Ploufragan, France

[2] Université européenne de Bretagne - Anses, Laboratoire de Ploufragan-Plouzané - Unité d'épidémiologie et bien-être aviaire, BP 53, Technopole Saint Brieuc Armor, 22440 Ploufragan, France

[3] Cnrs - Université de Lyon 1 - Unité de biométrie et biologie évolutive, UMR CNRS 5558, 43 bd du 11 novembre 1918, 69622 Villeurbanne, France

I - INTRODUCTION

1. BESOINS MÉTHODOLOGIQUES DES ÉPIDÉMIOLOGISTES

Le traitement statistique des données d'épidémiologie analytique animale a usuellement pour objectif de déterminer les facteurs de risque d'une maladie ou d'un problème de santé publique vétérinaire. Pour traiter ce type de données, les modèles linéaires généralisés, dont la régression logistique est un cas particulier, sont classiquement utilisés [Agresti, 2002 ; Dohoo *et al.*, 2010]. Pourtant, ces modèles présentent des limites, notamment lorsque les facteurs de risque potentiels sont nombreux et liés entre eux, cas fréquemment rencontré en pratique. Pour une discussion référencée sur ce sujet, le lecteur intéressé peut se référer à [Bougeard *et al.*, 2008 ; Bougeard, 2010]. Les limites associées aux modèles linéaires généralisés, ainsi que leurs nombreux avantages, servent de base au développement de nouvelles méthodes statistiques dont les objectifs sont une meilleure adaptation aux spécificités des données d'épidémiologie animale. Le cahier des charges de ces nouvelles méthodes est le suivant :

Obj. 1. Prendre en compte de nombreux facteurs de risque potentiels dans un même modèle. Ceci doit aider à interpréter de façon conjointe les facteurs de risque liés entre eux. Ce modèle doit aussi permettre la sélection automatique et non ambiguë des nombreux facteurs de risque ;

Obj. 2. Prendre en compte une maladie ou un problème de santé publique vétérinaire décrit par plusieurs variables, *e.g.*, lésions observées et résultats sérologiques, ou atteintes des différentes classes d'âges ;

Obj. 3. Bénéficier d'un modèle stable même en cas de colinéarités marquées entre les facteurs de risque potentiels. Développer des indices de qualité de ce modèle ;

Obj. 4. Tenir compte de la nature des variables, *i.e.*, quantitative ou qualitative ;

Obj. 5. Etre à la fois descriptive (description des liens entre les variables) et explicative (obtention d'un modèle liant facteur de risque et maladie). Ce modèle doit permettre d'obtenir des odds ratio pour quantifier les liens entre facteurs de risque et maladie ;

Obj. 6. Tenir compte de la structure des facteurs de risque potentiels en multiples tableaux (*e.g.*, variables relatives à la structure de l'élevage, à l'alimentation des animaux, à l'hygiène), cette structure ayant une signification zootechnique intéressante pour l'interprétation. Développer des indices d'aide à l'interprétation adéquats.

Deux caractéristiques n'ont pas été retenues dans un premier temps, *i.e.*, la prise en compte de la structure hiérarchisée des observations (*e.g.*, les animaux étudiés sont emboîtés dans des élevages) et dans une moindre mesure, la prise en compte de données manquantes. Ces deux points seront évoqués dans les perspectives.

2. DÉMARCHE ADOPTÉE

Afin de développer un outil original utile aux épidémiologistes, ces derniers sont directement impliqués dans chacune des étapes du développement. La démarche générale est subdivisée en plusieurs étapes. (1) La première étape consiste en l'expression des limites des méthodes standards et des besoins des épidémiologistes. Ces éléments sont détaillés dans la partie précédente I.1. Cette étape est cruciale car elle oriente les recherches méthodologiques ultérieures et assure que la méthode réponde réellement aux besoins exprimés. (2) La deuxième étape, dévolue aux seuls statisticiens, vise au développement méthodologique, ici de méthodes de régression multibloc à variables latentes dont le principe est expliqué dans la partie II suivante. Il faut noter que cette étape est généralement longue car elle nécessite la publication scientifique des méthodes afin qu'elles soient validées par d'autres statisticiens. De plus, ces nouvelles méthodes doivent être comparées aux méthodes alternatives existantes, dans notre cas principalement la régression PLS multibloc [Wold, 1984 ; Wangen et Kowalski, 1988] et l'analyse canonique généralisée avec un tableau de référence [Kissita, 2003], afin d'en démonter l'originalité [Bougeard *et al.*, 2006 ; Bougeard *et al.*, 2007b ; Bougeard *et al.*, 2007a]. Par ailleurs, des études de simulations sont menées pour illustrer leur propriétés, *e.g.*, robustesse au faible nombre d'individus, au grand nombre de variables, à la multicolinéarité entre celles-ci [Bougeard et Qannari, 2011 ; Bougeard *et al.*, 2011b]. Il faut noter que généralement, plusieurs méthodes répondant au cahier des charges peuvent être développées, publiées et ne pas s'avérer

intéressantes par la suite lors du traitement à grande échelle des données d'épidémiologie animale (*cf.* étape 4). (3) La troisième étape vise à développer des indices ainsi que des graphes d'aide à l'interprétation de ces méthodes, en lien avec les épidémiologistes [Bougeard *et al.*, 2011a ; Bougeard *et al.*, 2011c]. En effet, une méthode statistique, aussi intéressante soit-elle, doit être d'accès aisé pour les utilisateurs. (4) La quatrième étape est axée sur l'application de ces nouvelles méthodes et outils à de nombreux jeux de données. Même si les méthodes statistiques développées sont fondées sur une idée théorique intéressante, même si leurs propriétés sont illustrées sur la base de simulations, dans la pratique, certaines apparaissent plus intéressantes que d'autres pour l'interprétation. Il est crucial que cette étape de test soit réalisée conjointement avec de nombreux utilisateurs ainsi que sur des données les plus variées possibles [Bougeard *et al.*, 2008 ; Lupo *et al.*, 2010 ; Bougeard *et al.*, 2012 ; Bougeard et Cardinal, sous presse]. (5) La dernière étape vise au développement d'un support logiciel pour mettre aisément ces outils à disposition de tous. En effet, aussi intéressante soit une méthode, si elle n'est pas disponible facilement, elle ne sera pas utilisée. Depuis une dizaine d'années, les nouvelles méthodes développées sont généralement diffusées aux autres statisticiens ainsi qu'aux utilisateurs potentiels grâce à des packages développés sur le logiciel libre R [Logiciel R, 2008]. Dans cette démarche, nous avons choisi, en plus, d'intégrer les nouvelles méthodes au logiciel ade4, développé sous R, afin de les associer aux nombreuses possibilités graphiques existantes et de permettre aux utilisateurs l'accès à d'autres méthodes d'intérêt [Dray et Dufour, 2007 ; Dray *et al.*, 2007].

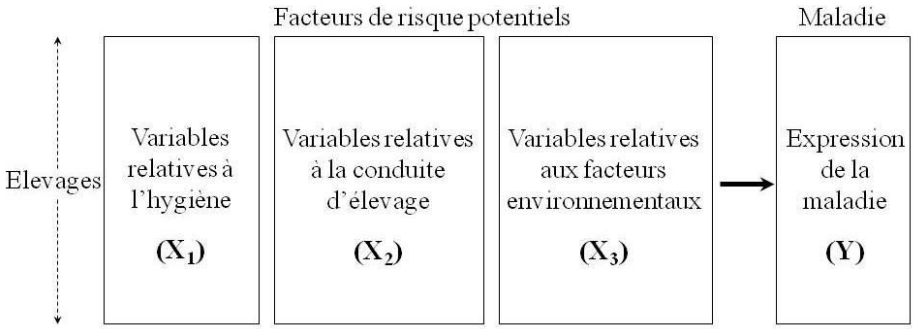
II - MÉTHODE

1. OBJECTIFS

A la vue du cahier des charges défini (*cf*. partie I.1), les méthodes de régression multibloc à variables latentes apparaissent comme une alternative prometteuse aux modèles linéaires généralisés. En effet, les méthodes multiblocs sont des extensions des méthodes d'analyse factorielle pour le cas des données organisées en plusieurs tableaux ($\mathbf{X}_1$, …, $\mathbf{X}_K$). Les méthodes de régression multibloc sont des extensions supervisées de ces dernières, c'est-à-dire que la description des données organisées en plusieurs tableaux ($\mathbf{X}_1$, …, $\mathbf{X}_K$) est orientée vers la description d'un tableau supplémentaire **Y**. Pour le cas des données d'épidémiologie animale, il s'agit d'expliquer une maladie décrite par plusieurs variables par de nombreux facteurs de risque potentiels organisés en multiples tableaux. Un exemple de ce type de données est proposé par la figure 1.

Figure 7

Illustration de données d'épidémiologie animale structurées en blocs de variables



Ces méthodes permettent de prendre en compte directement les objectifs 1 (nombreuses variables explicatives), 2 (plusieurs variables à expliquer), 5 (description et prédiction) et 6 (variables explicatives structurées en plusieurs blocs) définis dans la partie I.1. Il reste donc à les adapter pour qu'elles puissent s'ajuster aux objectifs 3 (colinéarité) et 4 (variables de différentes natures). Par ailleurs, ces méthodes sont relativement nouvelles et donc encore assez peu utilisées en pratique. Leurs propriétés doivent donc être mieux connues et des aides à l'interprétation spécifiques doivent être développées. De plus, ces méthodes n'ont pas été encore appliquées en épidémiologie animale et des ajustements devront être réalisés pour assurer leur bonne adéquation.

2. MÉTHODES DE RÉGRESSION MULTIBLOCS

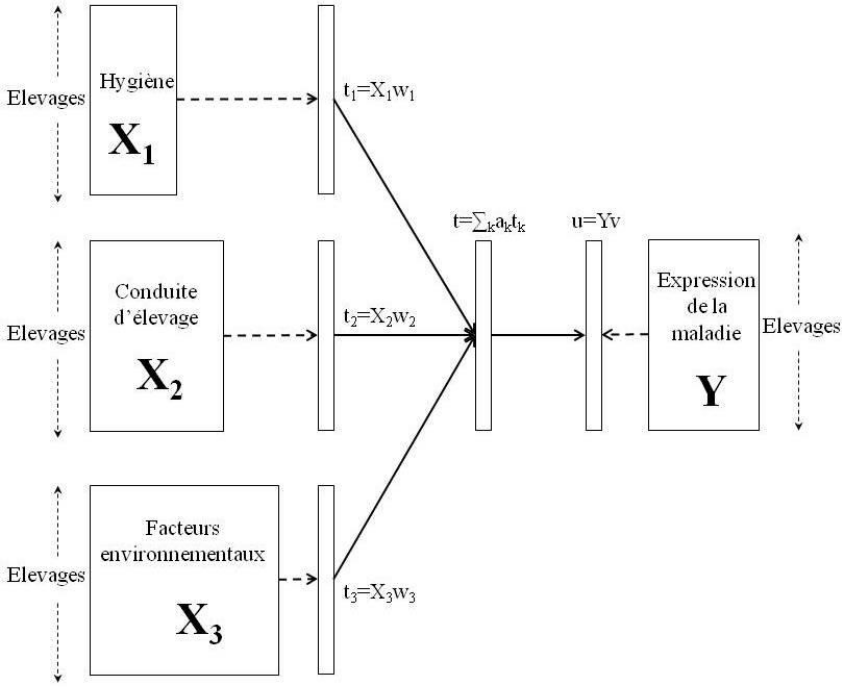
S'inspirant des idées des modèles à équations structurelles, dont LISREL [Joreskog, 1970] et l'approche PLS [Wold, 1982] sont les méthodes plus connues, les méthodes de régression multiblocs sont fondées sur la conjonction d'un modèle interne où chaque bloc de variables est résumé par une variable latente (encore appelée composante) et d'un modèle externe où les variables latentes sont liées entre elles par un modèle défini par l'utilisateur. Parmi les méthodes de régression multiblocs existantes, seules deux présentant un réel intérêt en terme méthodologique et pour l'interprétation, sont détaillées par la suite : la régression PLS multiblocs [Wold, 1984] et l'analyse des redondances multiblocs encore appelée ACPVI multiblocs [Bougeard *et al.*, 2007a]. Pour la justification de ce choix, le lecteur intéressé peut se référer à [Bougeard, 2010].

Le modèle interne consiste à rechercher le meilleur résumé de chaque tableau, soit pour chaque tableau explicatif (facteurs de risque potentiels) $t_k=\mathbf{X}_k w_k$ pour k=(1, .., K) mais aussi pour le tableau à expliquer (maladie) $u=\mathbf{Y}v$, les vecteurs w_k et v étant les axes respectivement associés aux variables latentes t_k et u. Selon l'objectif de l'analyse, *i.e.*, méthode explicative de la maladie mais sensible aux variables corrélées, ou méthode moins explicative mais plus stable, des contraintes de normes spécifiques peuvent être choisies. Pour le premier cas, correspondant à l'analyse des redondances multiblocs, les contraintes choisies sont $||t_k||=||v||=1$. Pour le second cas qui correspond à la régression PLS multiblocs, les contraintes sont $||w_k||=||v||=1$. Le modèle externe cherche à maximiser le lien entre ces variables selon la formule $\sum_k \text{cov}^2(t_k, u)$, « cov » désignant la covariance. Dans le but de représenter toutes les variables dans un même espace, mais aussi de prédire le tableau à expliquer **Y** par l'ensemble des facteurs de risque potentiels $\mathbf{X}=[\mathbf{X}_1|\ldots|\mathbf{X}_K]$, une variable latente $t=\mathbf{X}w$ est aussi recherchée selon la formule $t=\sum_k a_k t_k$. Ces variables latentes, ainsi que les liens les unissant, sont illustrées par la figure 2.

L'ensemble de ces éléments est synthétisé dans un critère à maximiser qui résout simultanément les modèles interne et externe. La solution de ce critère est issue de la décomposition en valeurs propres de la matrice $\sum_k \mathbf{X}_k'\mathbf{Y}(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{Y}'\mathbf{X}_k$ (analyse des redondances multiblocs) qui implique tableaux explicatifs $\mathbf{X}_k$ et tableau à expliquer **Y**, la structure en blocs des variables étant ainsi prise en compte. La solution de la régression PLS multiblocs est plus stable mais moins intéressante car fondée sur la décomposition de la matrice $\sum_k \mathbf{X}_k'\mathbf{Y}\mathbf{Y}'\mathbf{X}_k=\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$, ce qui revient à la solution de la régression PLS standard où la structure en blocs des variables n'est pas prise en compte. Pour tous les détails de résolution, le lecteur intéressé peut se référer à [Bougeard *et al.*, 2011b]. Cette solution fournit une variable latente pour chacun des tableaux, *i.e.*, $(t_1^{(1)}, \ldots, t_K^{(1)})$, $t^{(1)}$ et $u^{(1)}$ respectivement pour $(\mathbf{X}_1, \ldots, \mathbf{X}_K)$, **X** et **Y**. Cependant, les modèles biologiques étant rarement unidimensionnels, il est préférable de résumer les tableaux par plusieurs variables latentes, chacune apportant une information complémentaire à la précédente. Pour cela, l'information préalablement fournie par la première variable latente est soustraite à chacun des tableaux selon une procédure de déflation décrite par $\mathbf{X}_k=[\mathbf{I}-(t^{(1)}t^{(1)\prime})/(t^{(1)\prime}t^{(1)})]\mathbf{X}_k$ pour k=(1, …, K), **I** étant la matrice identité. Le critère à maximiser est ensuite appliqué à ces nouveaux tableaux déflatés pour obtenir les variables latentes de dimension supérieure $(t_1^{(h)}, \ldots, t_K^{(h)})$, $t^{(h)}$ et $u^{(h)}$ pour h=(2, …, H), avec H la dimension maximale du modèle qui correspond généralement au rang de la matrice explicative **X**.

Figure 2

Illustration du principe des méthodes de régression multiblocs à variables latentes



3. AIDES À L'INTERPRÉTATION

Les méthodes de régression multiblocs ont des objectifs à la fois descriptif et explicatif. L'interprétation descriptive est facilitée par des outils de visualisation graphique issus de l'analyse factorielle dont les méthodes de régression multiblocs sont des extensions. Les composantes globales t permettent de représenter l'ensemble des individus (*e.g.*, élevages ou lots d'animaux) dans un espace commun pour en appréhender les différences et similitudes. Par ailleurs, le graphe associé des liens entre variables est fondé sur les poids w* définis par t=**X**w* pour variables explicatives, et les coefficients c=(t't)$^{-1}$**Y**'t pour les variables à expliquer. La représentation factorielle de ces poids permet d'étudier les liens entre les variables appartenant à chaque bloc, explicatifs mais aussi à expliquer, et ainsi d'interpréter leurs corrélations et interactions.

L'interprétation explicative est fondée sur un modèle dont il convient dans un premier temps de sélectionner la dimension optimale (h.opt). En effet, H modèles peuvent être construits, fondés sur une, deux, .., ou H composantes. Pour cela, une validation croisée est utilisée afin de déterminer pour chacun de ces H modèles leur potentiel explicatif et prédictif [Stone, 1974]. Grâce à ces informations, un modèle optimal unique est sélectionné pour être interprété. Ce modèle permet de déterminer directement les coefficients de régression liant l'ensemble des variables à expliquer **Y** à l'ensemble des facteurs de risque potentiels **X** selon le modèle $\mathbf{Y}=\mathbf{X}[w^{*(1)}c^{(1)'}+\ldots+w^{*(h.opt)}c^{(h.opt)'}]+\varepsilon$ avec ε le résidu du modèle [Wold *et al.*, 1983]. Ce modèle étant fondé sur les composantes globales t telles que $\mathbf{Y}=(t^{(1)}c^{(1)'}+\ldots+t^{(h.opt)}c^{(h.opt)'})+\varepsilon$ et $t^{(h)}=\mathbf{X}w^{*(h)}$, orthogonales entre elles et résumées de toutes les variables explicatives, il apparaît que celui-ci est stable même lorsque les variables sont nombreuses et quasi-colinéaires. Ces coefficients de régression sont intéressants, d'autant plus qu'ils permettent le calcul des odds ratio, mais ils ne sont pas suffisants pour aider à l'interprétation des facteurs de risque lorsque la maladie **Y** est décrite par plusieurs variables. Pour cela, l'indice *VarImp* est développé pour comparer l'importance de tous les facteurs de risque potentiels dans l'explication globale de la maladie. Cet indice est calculé comme la somme des poids w*, pondérée par le poids de chaque bloc a_k. Il est mesuré sous forme de

pourcentage et est associé à un intervalle de confiance issu de simulations bootstrap ; la significativité de chaque variable explicative peut ainsi être définie [Freedman, 1981 ; Gosselin et al., 2010]. Dans le but d'obtenir une interprétation synthétique, l'indice *BlockImp* est développé pour mesurer l'importance de chaque bloc dans l'explication globale de la maladie. Cet indice est calculé par le poids de chaque bloc a_k^2. Comme l'indice précédent, il est donné sous forme de pourcentage associé à un intervalle de confiance calculé par simulations bootstrap. Tous les détails sur ces indices sont donnés dans [Bougeard *et al.*, 2011c].

III - DIFFUSION ET APPLICATION

1. DONNÉES D'ÉPIDEMIOLOGIE ANIMALE

Une illustration est proposée, à la fois de la méthode mais aussi du package R. Les données proviennent d'une enquête analytique menée en 2005 sur une cohorte de 351 lots de poulets de chair. L'objectif est d'identifier les facteurs de risque 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 vingt variables explicatives sélectionnées sont organisées en (K=4) tableaux, *i.e.*, $\mathbf{X}_1$ relatif à la structure de l'élevage (5 variables), $\mathbf{X}_2$ aux caractéristiques du lot la première semaine (4 variables), $\mathbf{X}_3$ aux caractéristiques du lot durant le reste de l'élevage (6 variables) et $\mathbf{X}_4$ au ramassage, conditions de transport et d'abattage (5 variables). Pour plus de détails sur les données, le lecteur intéressé peut se référer à [Lupo *et al.*, 2009]. Les variables catégorielles sont codées selon un codage disjonctif complet, pratique usuelle en analyse de données mixtes [Lebart *et al.*, 2000]. Comme les variables ont des unités de mesure différentes, elles sont centrées et réduites. Le descriptif des variables est donné par le tableau 1.

La finalité pour l'épidémiologiste est d'abord de comprendre les liens entre l'ensemble des variables, ainsi que les ressemblances et différences entre les lots étudiés. Puis, il apparaît primordial de déterminer les facteurs de risque des pertes globales (**Y**) et de façon plus précise de chaque élément constitutif de ces pertes (Y.Mort7, Y.Mort, Y.Doa, Y.Condemn). Au final, il est intéressant de bénéficier d'une vision globale des actions à mener dans les différentes phases de production ($\mathbf{X}_1$, …, $\mathbf{X}_4$) dans le but de réduire les pertes (**Y**).

Les données sont importées sur le logiciel R à partir du code suivant :

```
R> library(ade4)
R> data(chickenk)
R> Y <- chickenk[[1]]
R> Ychick <- dudi.pca(Y, scannf=F)
R> Xchick <- chickenk[2, 5]
```

Les données « chickenk » sont structurées en cinq tableaux (**Y**, $\mathbf{X}_1$, …, $\mathbf{X}_4$) répartis en un tableau à expliquer « Ychick » contenant 351 lignes (lots de poulets) et 4 variables (à expliquer), et un tableau explicatif « Xchick » contenant 351 lignes et 20 variables (facteurs de risque potentiels).

2. INTERPRÉTATION DESCRIPTIVE

Le package R développé propose deux méthodes aux utilisateurs, dont le choix dépend du degré de quasi-colinéarité des variables au sein de chaque bloc : la régression PLS multiblocs (fonction « mbpls ») et l'analyse des redondances multiblocs (fonction « mbpcaiv »). Les résultats de la régression multiblocs ainsi que les aides à l'interprétation descriptives associées sont donnés par le code R suivant :

```
R> Reschick <- mbpcaiv(Ychick, Xchick, scale=TRUE, option="uniform")

R> SummmaryChick <- summary(Reschick)

R> plot(Reschick)
```

La fonction « summary » permet de mesurer l'information issue de chacun des tableaux (**Y**, **X**, $\mathbf{X}_1$, .., $\mathbf{X}_4$) résumée par les dix premières dimensions (résultats non fournis ici). Il apparaît que les quatre premières variables latentes globales t permettent d'expliquer 90,6 % de l'information contenue dans les données.

Tableau 1

Description des variables du jeu de données relatif aux pertes en élevage de poulets de chair

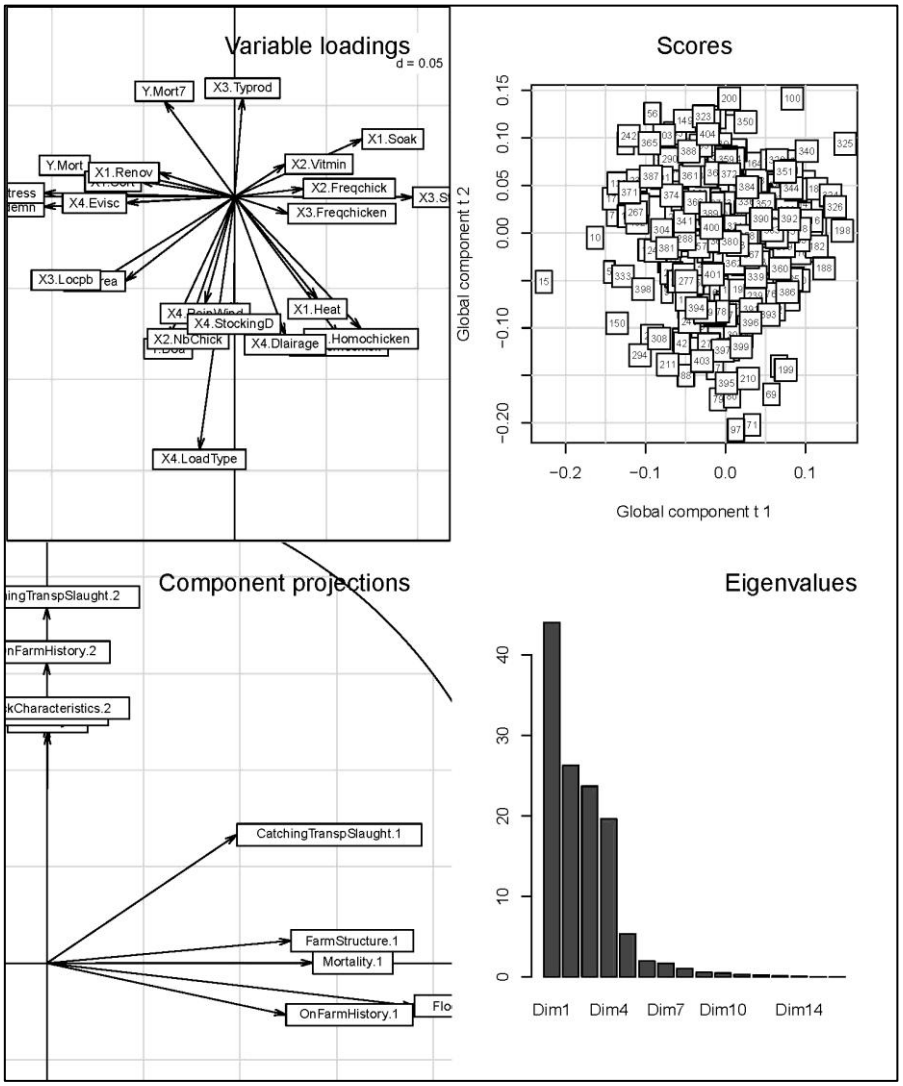
Bloc	Variable	Description de la variable
Y (Pertes)	Y.Mort7	Mortalité des poussins sept jours après la mise en place (%)
	Y.Mort	Mortalité des poulets deux jours avant l'abattage (%)
	Y.Doa	Mortalité des poulets pendant le ramassage et le transport (%)
	Y.Condemn	Taux de saisie des poulets à l'abattoir (%)
X_1 (Structure)	X1.Area	Surface totale du bâtiment de poulets (en m²)
	X1.Soak	Détrempage du bâtiment (non / oui)
	X1.Heat	Matériel de chauffage (radiant / canon)
	X1.Sort	Tri des animaux (non / oui)
	X1.Renov	Age du bâtiment (<12 ans ou rénové / >12 ans)
X_2 (Démarrage)	X2.Vitmin	Vitamines et minéraux au démarrage (non / oui)
	X2.Freqchick	Nombre de passages quotidiens au démarrage
	X2.Homochick	Homogénéité du lot de poussins à la mise en place (non /oui)
	X2.NbChick	Nombre de poussins à la mise en place
X_3 (Croissance)	X3.Typrod	Type de lot (standard / autre)
	X3.Homochicken	Homogénéité du lot au cours de l'élevage (non / oui)
	X3.Strain	Souche génétique (A / B)
	X3.Locpb	Problème locomoteur (non / oui)
	X3.Stress	Présence d'un stress lors de l'élevage des animaux (*e.g.*, problème d'alimentation, de chauffage)
	X3.Freqchicken	Nombre de passages de l'éleveur pendant la période d'élevage (nombre/jour)
X_4 (Ramassage, abattage)	X4.LoadType	Mode de mise en caisse (manuel / mécanique)
	X4.RainWind	Pluie et vent lors de l'attente des animaux (non / oui)
	X4.StockingD	Densité des animaux dans les caisses de transport (en kg/m²)
	X4.Dlairage	Durée moyenne d'attente sur le quai (en minutes)
	X4.Evisc	Présence d'un opérateur de retrait à l'éviscération (non / oui)

Ces variables latentes expliquent 92,6 % de la variance du tableau **Y** et 31,5 % du tableau **X**, ce qui est logique au vu du nombre de variables de chaque tableau et des objectifs de l'analyse. De plus, ces variables latentes t expliquent 19,5 % de la variance de $\mathbf{X}_1$ (structure de l'élevage), 34,7 % de $\mathbf{X}_2$ (caractéristiques du lot la première semaine), 35,4 % de $\mathbf{X}_3$ (caractéristiques du lot durant le reste de l'élevage) et 35,6 % de $\mathbf{X}_4$ (ramassage, conditions de transport et d'abattage).

La fonction « plot » donne accès aux graphiques de la figure 3. Le premier (Variable loadings) permet d'étudier les liens entre toutes les variables. Il apparaît que les quatre éléments caractérisant la mortalité sont peu liés entre eux et que chacun est lié à des facteurs de risque spécifiques. La mortalité durant la première semaine (Y.Mort7) par exemple, est négativement liée à l'homogénéité des poussins à la mise en place (X2.Homochick) et au système de chauffage par canons (X1.Heat). Il faut noter que l'homogénéité du lot au cours de l'élevage (X3.Homochicken) est corrélée à l'homogénéité des poussins à la mise en place (X2.Homochick). Le deuxième graphe (Scores) donne le positionnement de chacun des 351 lots de poulets au regard de ces variables. Le troisième graphe (Component projections) permet de visualiser le lien entre chaque composante partielle (t_1, …, t_4) et la composante globale t pour les deux premières dimensions, ce qui illustre les liens entre blocs.

Figure 3

Aide à l'interprétation descriptive de l'analyse des redondances multiblocs appliquée aux données de pertes en élevages de poulets de chair



Il s'ensuit que la première composante associée aux pertes « Mortality1 », plutôt associée à la mortalité durant l'élevage (Y.Mort) et aux saisies à l'abattoir (Y.Condemn), est liée à la structure de l'élevage ($\mathbf{X}_1$), aux caractéristiques du lot la première semaine ($\mathbf{X}_2$) et durant le reste de l'élevage ($\mathbf{X}_3$). Le dernier graphe (Eigenvalues) reprend l'information synthétisée par chaque composante.

3. INTERPRÉTATION DU MODÈLE MULTIBLOCS

La première étape consiste à sélectionner le modèle optimal par validation croisée en suivant le code R suivant :

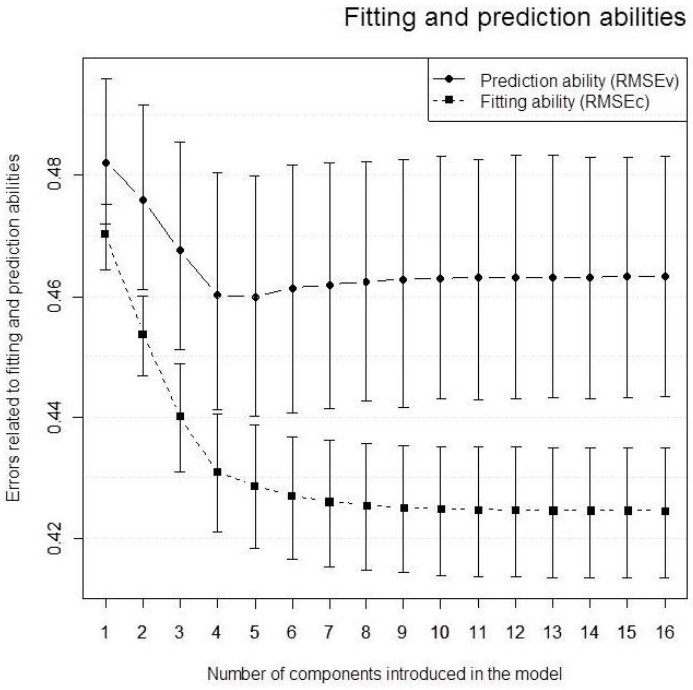
```
R> TestdimChick <- tesdim.multiblock(Reschick, nbcross=500)

R> plot(TestdimChick)
```

Cinq cents ré-échantillonnages du jeu de données sont constitués. L'ensemble de ces résultats est synthétisé par la fonction « plot » qui donne accès à la figure 4. Le modèle ayant la dimension optimale est celui qui minimise à la fois l'erreur de modélisation (RMSEc) et l'erreur de prédiction (RMSEv). Un modèle à quatre dimensions est ici sélectionné.

Figure 4

Capacité de modélisation (RMSEc) et de prédiction (RMSEv) du modèle issu de l'analyse des redondances multiblocs appliquée aux données de pertes en élevages de poulets de chair



Dans un deuxième temps, il est important de pouvoir évaluer la qualité du modèle optimal. Pour cela, la fonction « plot.modelquality.multiblock » donne accès aux graphes de diagnostic donnés figure 5.

```
R> plot.modelquality.multiblock(Reschick,
SummmaryChick, dimopt=4)
```

Le premier graphe reprend la variance expliquée de chaque tableau par le modèle optimal, ainsi que les coefficients de détermination (R^2 et R^2 ajusté). La deuxième ligne de graphes (Observed versus fitted values) représente, pour chaque variable à expliquer, les valeurs observées *versus* prédites des 351 observations. Il est intéressant de constater que les variables Y.Mort7 et Y.Mort sont bien modélisées alors que la variable Y.Condemn l'est moins bien. La troisième ligne de graphes (Residual scaterring) contrôle l'indépendance des résidus du modèle en constatant (ou pas) leur répartition aléatoire autour de la valeur zéro ; les observations sur- ou sous-estimées sont repérées. Pour chacune des variables à expliquer, la dispersion des observations est plus grande pour les valeurs prédites élevées relativement aux faibles, ce qui est dû à la distribution asymétrique de celles-ci.

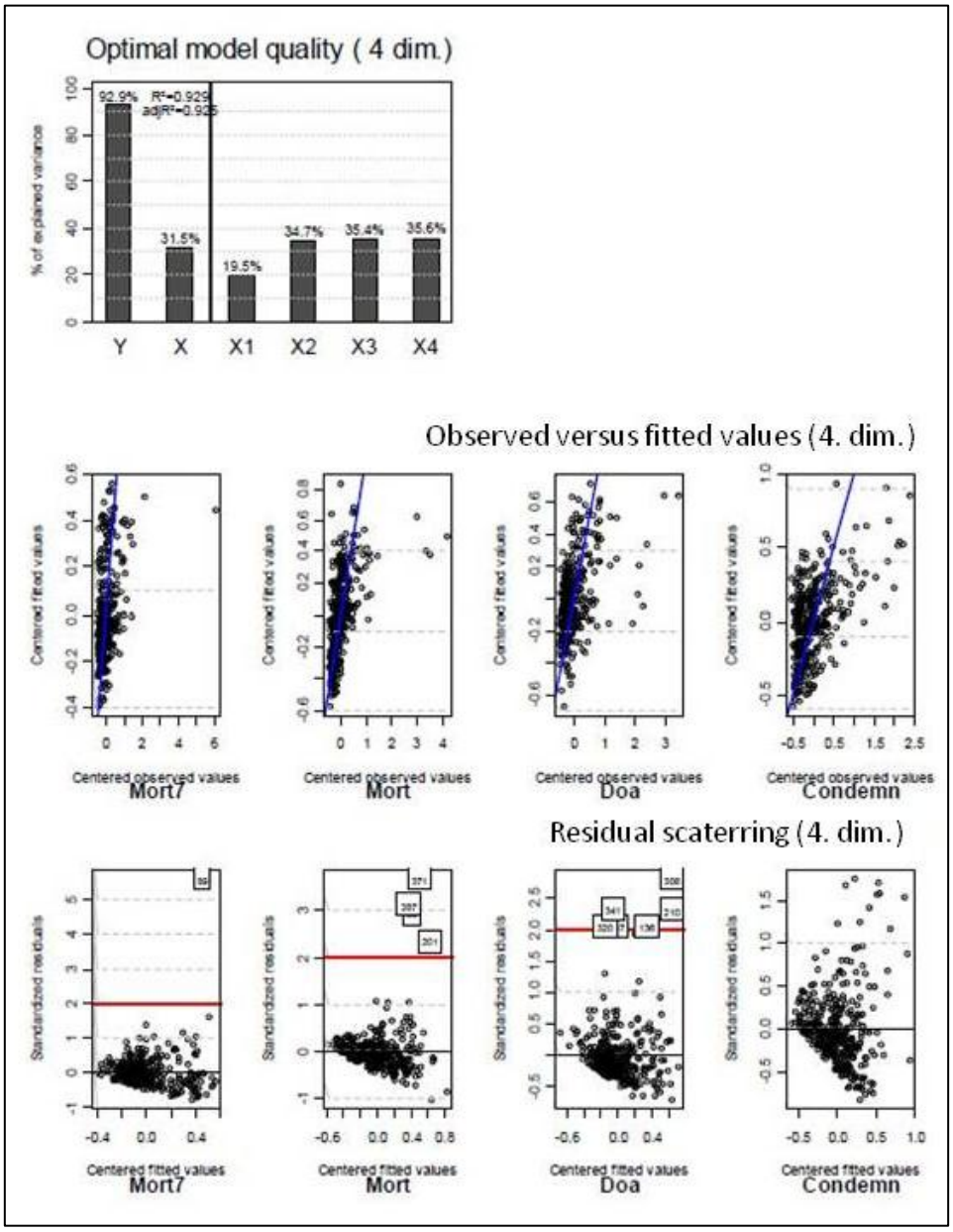
Finalement, les simulations bootstrap donnent accès à des intervalles de confiance associés aux paramètres explicatifs du modèle optimal, *i.e.*, coefficients de régression (première ligne de graphe), importance des variables dans l'explication des pertes (indice *VarImp*, second graphe) et importance des blocs de variables dans l'explication des pertes (indice *BlockImp*, troisième graphe). Ces résultats sont illustrés par la figure 6. Il faut noter que les coefficients de régression et leurs intervalles de confiance sont aisément transformables en odds ratios. La fonction « plot » illustre ces indices d'aide à l'interprétation explicative.

```
R> BootChick <- bootstrap.multiblock(Reschick,
nboot=500, dimopt=4)
```

```
R> plot(BootChick)
```

Figure 5

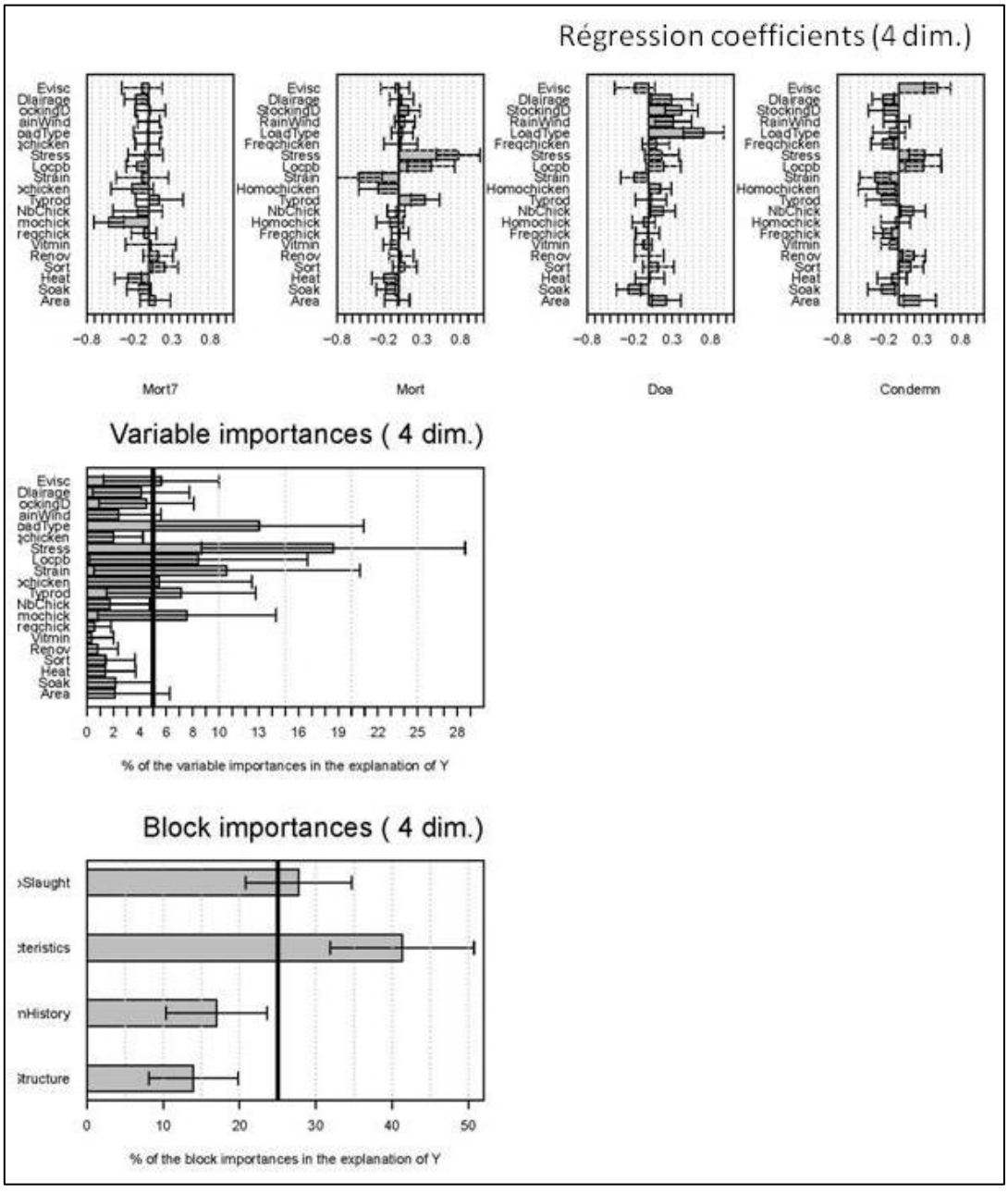
Indices de qualité du modèle optimal issu de l'analyse des redondances multiblocs appliquée aux données de pertes en élevages de poulets de chair



Il s'ensuit que la mortalité des poulets avant l'abattage (Y.Mort) est significativement liée à la présence d'un stress lors de l'élevage, *e.g.*, panne électrique, de ventilation ou d'alimentation (X3.Stress), aux problèmes locomoteurs (X3.Locpb) et à la souche génétique (X3.Strain). La mortalité des poulets pendant le ramassage et le transport (Y.Doa) est plutôt associée au mode de ramassage (X4.LoadType), à la densité des animaux dans les caisses de transport (X4.StockingD) et à la présence de pluie et vent lors de l'attente des animaux (X4.RainWind). De façon plus globale, les pertes en élevage de poulet de chair sont significativement liées aux stress des animaux durant la période d'élevage, ainsi qu'au type de ramassage des animaux (mécanique *versus* manuel). Au final, il apparaît que le bloc relatif aux caractéristiques du lot durant l'élevage ($\mathbf{X}_3$) est celui qui a la plus forte influence sur les pertes.

Figure 6

Aide à l'interprétation explicative de l'analyse des redondances multiblocs appliquée aux données de pertes en élevages de poulets de chair



IV - CONCLUSION ET PERSPECTIVES

Afin de répondre aux besoins statistiques des épidémiologistes, issus notamment des limites des modèles linéaires généralisés, des méthodes de régression multiblocs à variables latentes sont étudiées et rendues accessibles. Ces méthodes permettent de prendre en compte de nombreux facteurs de risque potentiels organisés en blocs de variables ayant du sens en termes d'interprétation ; de plus, ces méthodes permettent d'expliquer une maladie pouvant être décrite par plusieurs variables. Plusieurs méthodes multiblocs ont été évaluées sur la base de leurs propriétés théoriques,

de simulations de leur comportement selon le format des données, ainsi que par le traitement de nombreux jeux de données. Deux méthodes ont été retenues puis développées dans un package R, libre d'accès pour tout utilisateur potentiel, statisticien, épidémiologiste ou autre. Ont été associés à ce package des outils d'aide à l'interprétation, d'ordre descriptif et prédictif, répondant aux besoins des épidémiologistes. Les méthodes multiblocs permettent d'extraire une grande quantité d'information issue de données d'épidémiologie complexes, mais aussi de la synthétiser sous forme d'indices et de graphes.

Afin de tenir compte de données d'épidémiologie plus complexes, de nouveaux développements sont en cours ou à prévoir. Par exemple, la prise en compte de liens non linéaires entre variables explicatives et à expliquer permettrait d'améliorer les modèles, selon les idées proposées par [Verdun *et al.*, 2012]. Par ailleurs, la prise en compte d'observations structurées en groupes est un sujet de recherche actif en analyse multiblocs et fait l'objet d'un travail de thèse récent [Eslami, 2013]. De plus, dans ce travail, des algorithmes NIPALS ont été développés pour chaque méthode proposée (mais non encore évalués), ce qui devrait permettre leur utilisation prochaine pour le cas de données manquantes [Wold, 1975]. La démarche présentée ici sera appliquée à ces recherches afin de mettre à disposition des épidémiologistes un outil nouveau et performant pour des données ayant des variables organisées en blocs et présentant une structure en groupe des observations.

BIBLIOGRAPHIE

Agresti A. - Categorical data analysis, 2ième édition Ed, Hoboken, New Jersey, 2002.

Bougeard S. - Description et prédiction à partir de données structurées en plusieurs tableaux. Application en épidémiologie animale. Editions Universitaires européennes. 2010. ISBN 978-613-1-52069-3.

Bougeard S., Cardinal M. - Multiblock modeling for complex preference study. Application to European preferences for smoked salmon. *Food Qual. Prefer.*, sous presse.

Bougeard S., Hanafi M., Noçairi H., Qannari E.M. - Multibloc canonical correlation analysis for categorical variables: application to epidemiological data (chap.17), *In:* Multiple correspondence analysis and related methodsChapman & Hall (Ed.), 2006, 393-404.

Bougeard S., Hanafi M., Qannari E.M. - ACPVI multibloc. Application à des données d'épidémiologie animale. *Journal de la Société Française de Statistique*, 2007a, **148**, 77-94.

Bougeard S., Hanafi M., Qannari E.M. - Multiblock latent root regression. Application to epidemiological data. *Computational statistics*, 2007b, **22**, 209-222.

Bougeard S., Lupo C., Le Bouquin S., Qannari E.M., Chauvin C. - Use of multiblock modelling to assess the overall risk factors of a composite outcome. Illustration on losses in broiler chickens. *Epidemiol. Infect.*, 2012, **140**, 337-347.

Bougeard S., Qannari E.M. - Continuum Approach for Multiblock Methods: Overview and Regularization Purpose. *14th Conference of the Applied Stochastic Models and Data Analysis International Society*, Roma, Italy, 2011.

Bougeard S., Qannari E.M., Hanafi M., Madec F., Rose N. - Proposition d'une méthode factorielle multibloc pour le traitement des données d'épidémiologie animale. *Épidémiologie et Santé Animale*, 2008, **53**, 1-10.

Bougeard S., Qannari E.M., Lupo C., Chauvin C. - Multiblock Redundancy Analysis from a user's perspective. Application in veterinary epidemiology. *Electronic Journal of Applied Statistical analysis and data mining*, 2011a, **4**, 203-214.

Bougeard S., Qannari E.M., Lupo C., Hanafi M. - From Multiblock Partial Least Squares to Multiblock Redundancy Analysis. A Continuum Approach. *Informatica*, 2011b, **22, 11-26.**

Bougeard S., Qannari E.M., Rose N. - Multiblock Redundancy Analysis : interpretation tools and application in epidemiology. *J. Chemometr.*, 2011c, **25**, 467-475.

Dohoo I.R., Martin W., Stryhn H. - Veterinary epidemiologic research (2nd edition), Prince Edward Island, Canada, 2010.

Dray S., Dufour A.B. - The ade4 Package: Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, 2007, **22**, 1-20.

Dray S., Dufour A.B., Chessel D. - The ade4 Package: II: Two-table and K-table Methods. *R News*, 2007, **7**, 47-52.

Eslami A. - Analyses factorielles de données structurées en groupes d'individus. Application en biologie (Thèse). Université de Rennes 1, 2013.

Freedman D.A. - Bootstrapping regression models. *Annals of Statistics*, 1981, **9**, 1218-1228.

Gosselin R., Rodrigue D., Duchesne C. - A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications. *Chemom. Intell. Lab.*, 2010, **100**, 12-21.

Joreskog K.G. - A general method for analysis of covariance structure. *Biometrika*, 1970, **57**, 239-251.

Kissita G. - Les analyses canoniques généralisées avec tableau de référence généralisé : éléments théoriques et appliqués (Thèse). Université de Paris Dauphine IX, 2003.

Lebart L., Morineau A., Piron M. - Statistique exploratoire multidimensionnelle, 3ième édition Ed, Paris, 2000.

Logiciel R version 3.0.1 - 2008. http://www.r-project.org/, The R project.

Lupo C., Bougeard S., Balaine L., Michel V., Petetin I., Colin P., Le Bouquin S., Chauvin C. - Risk factors for sanitary condemnation in broiler chickens and their relative impact. Application of an original multiblock approach. *Epidemiol. Infect.*, 2010, **138**, 364-365.

Lupo C., Le Bouquin S., Balaine L., Michel V., Peraste J., Petetin I., Colin P. et Chauvin C. - Feasibility of screening broiler chicken flocks for risk markers as an aid for meat inspection. *Epidemiol. Infect.*, 2009, **137**, 1086-1098.

Stone M. - Cross-validatory choice and assessment of statistical predictions. *J. Roy. Stat. Soc.*, 1974, **36**, 111-147.

Verdun S., Hanafi M., Cariou V., Qannari E.M. - Quadratic PLS1 regression revisited. *J. Chemometr.*, 2012, **26**, 384-389.

Wangen L.E., Kowalski B.R. - A multiblock partial least squares algorithm for investigating complex chemical systems. *J. Chemometr.*, 1988, **3**, 3-20.

Wold H. - Path models with latent variables. The NIPALS approach. , *In:* Quantitative Sociology: International perspectives on mathematical and statistical model building. H. M. Blalock A.A., F. M. Borodkin, R. Boudon, & V. Capecchi (Eds.), Editor(Ed.), 1975, 307-357.

Wold H. - Soft modelling: the basic design and some extensions, *In:* System under indirect observation. Part 2. K.G Jöreskog & Wold H., Editor(Ed.): North-Holland, Amsterdam, 1982, 1-54.

Wold S. - Three PLS algorithms according to SW. *Symposium MULDAST (multivariate analysis in science and technology)*, Umea University, Sweden, 1984.

Wold S., Martens H., Wold H. - The multivariate calibration problem in chemistry solved by the PLS method. *Proceedings of the Conference on Matrix Pencils*, 1983.