PROPOSITION D'UNE METHODE FACTORIELLE MULTIBLOC POUR LE TRAITEMENT DES DONNEES D'EPIDEMIOLOGIE ANIMALE *

Stéphanie Bougeard [1], **El Mostafa Qannari** [2], **Mohamed Hanafi** [2], **François Madec** [1] et **Nicolas Rose** [1]

RESUME

Le traitement statistique des enquêtes d'épidémiologie analytique animale est habituellement effectué à l'aide de méthodes qui s'apparentent aux modèles linéaires généralisés. Pourtant, le format des données pose souvent des problèmes relatifs à l'utilisation de ces modèles. La première limite est que les nombreuses variables explicatives ne peuvent pas toutes être intégrées au modèle. La seconde limite est que ces variables présentent des quasi-colinéarités marquées, ce qui affecte la pertinence et la stabilité des résultats. La troisième limite est relative à la structure des variables explicatives en blocs ayant une signification zootechnique intéressante dont il apparaît important de mesurer le poids dans l'explication de la maladie. Les phénomènes de santé animale étant complexes, la maladie est souvent décrite par plusieurs variables, ce qui constitue la dernière limite. Afin de résoudre ces difficultés, nous proposons une solution qui se positionne dans le cadre des méthodes factorielles multiblocs. Cette méthode, appelée analyse des redondances multibloc, permet la description et la prédiction de données structurées en plusieurs tableaux de variables explicatives orientés vers l'explication d'un autre tableau, la maladie. Les résultats illustrant cette méthode sont issus d'une enquête analytique visant à déterminer les facteurs de risque du niveau de séroprévalence intra-élevage au circovirus porcin de type 2.

Mots-clés : Epidémiologie analytique, facteur de risque, analyse factorielle multibloc, analyse des redondances, porc, sérologie PCV2.

SUMMARY

The analysis of animal epidemiological surveys is usually performed using methods related to generalized linear models. But some difficulties arise regarding the use of these models due to the nature of the data collected. The first limitation is that all the explanatory variables cannot be included in the model. The second limitation is the multicolinearity among the explanatory variables, which is likely to lead to a non-relevant and unstable model. The third limitation is that explanatory variables are organized in meaningful blocks. It seems of importance to assess which blocks of explanatory variables are good predictors of the disease. The last limitation for the use of generalized linear models is that the expression of the animal disease is often described by several variables. In order to circumvent these difficulties, we propose a new method in the field of multiblock factorial method. Multiblock redundancy analysis is a method to be used for the purpose of exploring and modeling the relationships of a set of several data tables, where we wish to predict a dataset (the disease) from several other datasets. The interest of the method is illustrated on the basis of a cross-sectional study in the field of veterinary epidemiology, which was carried out in order to assess the risk factors for seroprevalence of porcine circovirus type-2 in a pig population.

Keywords : Analytical epidemiology, Risk factors, Multiblock factorial analysis, Redundancy analysis, Pig, PCV2 serology.

* Texte de la communication orale présentée lors des Journées AEEMA, 22-23 mai 2008

[1] AFSSA, Département d'épidémiologie animale - Zoopôle, BP53, 22440 Ploufragan, France

[2] ENITIAA-INRA, Unité de sensométrie et chimiométrie - Rue de la Géraudière BP 82225, 44322 Nantes Cedex, France

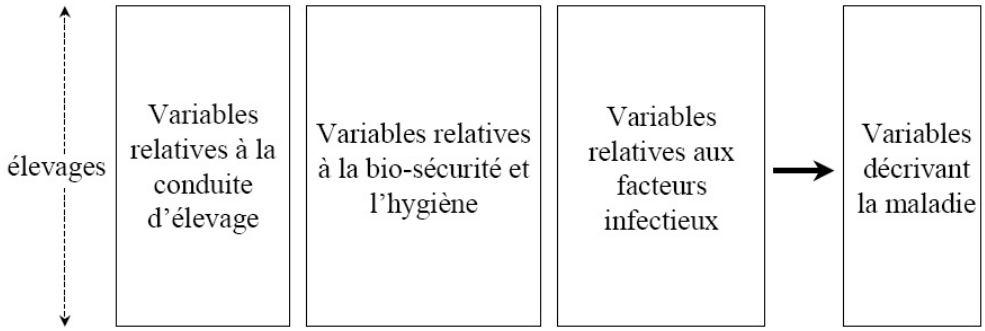
I - INTRODUCTION

1. CARACTERISTIQUES DES DONNEES D'EPIDEMIOLOGIE ANALYTIQUE

Dans le domaine vétérinaire, l'épidémiologie analytique est une discipline dédiée à l'étude des causes apparentes et des évènements directement ou indirectement associés à une maladie ou un phénomène de santé publique vétérinaire [Toma *et al.*, 1996]. Ces études peuvent être réalisées au travers d'enquêtes menées en élevage, à l'abattoir et dans des laboratoires de diagnostic. Que les données collectées soient issues d'une enquête de type cas/témoins ou exposés/non exposés, leur structure ainsi que les objectifs de traitement statistique associés sont comparables. Les variables explicatives, potentiellement facteurs de risque, sont généralement structurées en thèmes : les caractéristiques de la ferme (taille de l'élevage, performances zootechniques, autres productions animales, …), la conduite d'élevage (taux de renouvellement, technique de reproduction, nombre d'animaux par portée, …), l'habitat des animaux (mode de ventilation, isolation, chauffage, enregistrements bio-climatiques, …), l'alimentation et l'abreuvement des animaux (enregistrements alimentaires, mode de distribution, nombre de mangeoires, origine des aliments, …), l'état sanitaire du troupeau (dosages sérologiques, pesées, maladies chroniques, taux de réformes, vaccinations, traitements antibiotiques, …), les pratiques d'hygiène (protocoles de nettoyage et de désinfection) et les mesures de bio-sécurité (introduction d'animaux venant de l'extérieur, protection contre les nuisibles, …). La maladie ou le phénomène de santé étudié sont souvent complexes et décrits par plusieurs variables à expliquer (taux de mortalité, signes cliniques observés en élevage et lésions observées à l'abattoir, …). Un exemple de structure simplifiée de ces données est résumé par la figure 1. Il faut noter, de plus, que l'ensemble des données collectées est de nature mixte (*i.e.* mélange de variables qualitatives et quantitatives). Bien qu'il y ait une pré-sélection des variables explicatives, celles-ci restent nombreuses et structurellement liées entre elles. En effet, la multicolinéarité entre les variables explicatives n'est généralement significative que pour un nombre réduit de paires de variables, mais la quasi-totalité des variables est concernée. Il est souvent impossible de résoudre le problème de multicolinéarité en ne supprimant que quelques variables.

Figure 1

Structure usuelle des données d'épidémiologie analytique en santé animale



2. AVANTAGES ET LIMITES DES TRAITEMENTS STATISTIQUES USUELS

Le traitement statistique de ces données est classiquement basé sur une régression, qui permet d'expliquer la maladie par les variables explicatives présumées influentes. Les méthodes de régression les plus utilisées sont des cas particuliers des modèles linéaires généralisés [Agresti, 2002]. Une synthèse bibliographique du journal *Preventive Veterinary Medicine* montre que ce sont principalement la régression logistique (54% des cas utilisant un modèle linéaire généralisé), et dans une moindre mesure la

régression de Cox (23% des cas) et la régression linéaire (12% des cas), qui sont utilisées [Bougeard, 2007]. Ces méthodes présentent de nombreux avantages pour le traitement des données d'épidémiologie. Elles permettent tout d'abord de quantifier le lien entre une variable à expliquer (la maladie) et un ensemble de variables, potentiellement facteurs de risque, au travers de l'*odds ratio* ou du risque relatif. L'influence éventuelle d'interactions entre variables explicatives vis-à-vis de la maladie peut aussi être quantifiée. Les modèles linéaires généralisés permettent de plus de prendre en compte diverses natures de variables *(i.e.* qualitatives ou quantitatives). L'extension de ces modèles aux modèles linéaires généralisés mixtes permet d'intégrer d'éventuelles structures emboîtées des individus [Dohoo *et al.*, 2003].

Cependant, certains problèmes subsistent à l'utilisation de ces modèles. La première et plus importante limite est que les variables explicatives présentent des quasi-colinéarités marquées, ce qui affecte la pertinence et la stabilité des résultats issus de ces modèles [Schaeffer, 1986 ; Hosmer et Lemeshow, 1989 ; Weissfeld et Sereika, 1991 ; Dohoo *et al.*, 1997 ; Allison, 1999]. La seconde limite est que, en regard du nombre de degrés de liberté du modèle (souvent limité par le nombre d'individus), ces nombreuses variables, potentiellement facteurs de risque, ne peuvent pas toutes être intégrées au modèle. La troisième limite est relative à la structure des variables explicatives en blocs ayant une signification zootechnique intéressante (pratique d'hygiène, conduite d'élevage, …) dont il apparaît important de mesurer le poids dans l'explication de la maladie. Les phénomènes de santé animale étant complexes, la maladie est souvent décrite par plusieurs variables, ce qui constitue la dernière limite. En effet, l'épidémiologiste n'est satisfait ni par l'explication d'une variable de synthèse, ni par des modélisations disjointes.

3. DES SOLUTIONS POSSIBLES GRACE A L'ANALYSE FACTORIELLE

L'objectif est de décrire et modéliser les liens entre un grand nombre de variables explicatives, liées entre elles et organisées en groupes, et plusieurs variables à expliquer. Nous choisissons de nous situer dans le cadre général de l'analyse factorielle, propice à l'étude de ce type de données, son but étant de décrire un vaste ensemble de variables, inter-agissant entre elles, et dont on ne connaît *a priori* pas la structure. Pour cela, l'analyse factorielle décrit et synthétise l'information contenue dans le(s) tableau(x) de données grâce à des composantes (appelées aussi variables latentes), combinaisons linéaires des variables le(s) constituant. La sélection de variables, nécessaire et contraignante dans les techniques de régression usuelles, n'est plus indispensable ; l'information prise en compte est plus riche. Les résultats peuvent être donnés sous forme de représentations graphiques synthétiques, permettant d'appréhender les liens complexes unissant variables et/ou individus [Lebart *et al.*, 2000]. Ces méthodes sont assez peu utilisées par les épidémiologistes ; le nombre d'articles relatifs à leur utilisation dans le journal *Preventive Veterinary Medicine* représente 1% des articles parus, contre 20% relatifs à l'utilisation des modèles linéaires généralisés [Bougeard, 2007]. Pourtant, les solutions proposées par l'analyse factorielle, et notamment ses développements actuels, peuvent répondre aux limites des modèles linéaires généralisés dans le cadre du traitement des données d'épidémiologie animale.

II - PROPOSITION D'UNE METHODE FACTORIELLE MULTIBLOC

1. PRESENTATION GENERALE DE L'ANALYSE DES REDONDANCES MULTIBLOC

Nous proposons une solution se positionnant dans le récent cadre des méthodes factorielles multiblocs permettant la description, mais aussi la prédiction, de données structurées en plusieurs tableaux : plusieurs tableaux de variables explicatives orientés vers l'explication d'un autre tableau (la maladie). Des informations nouvelles sont ainsi apportées à l'épidémiologiste : il devient possible non seulement de décrire des liaisons entre variables explicatives, orientées vers l'explication de la maladie, mais aussi de mesurer l'influence de chaque bloc de variables explicatives dans l'explication de la maladie. Ces méthodes associent de plus

analyse factorielle et régression, étape indispensable du traitement des données d'épidémiologie animale. Dans la régression, les variables explicatives sont remplacées par les composantes issues de l'analyse factorielle. Ces techniques de régression orthogonalisée [Massy, 1965 ; Saporta, 1975 ; Lafi et Kaneene, 1992] diminuent nettement le problème de la multicolinéarité en utilisant des composantes orthogonales mutuellement, et en écartant de la régression les composantes considérées comme étant liées au bruit [Lebart *et al.*, 2000 ; Barker et Brown, 2001].

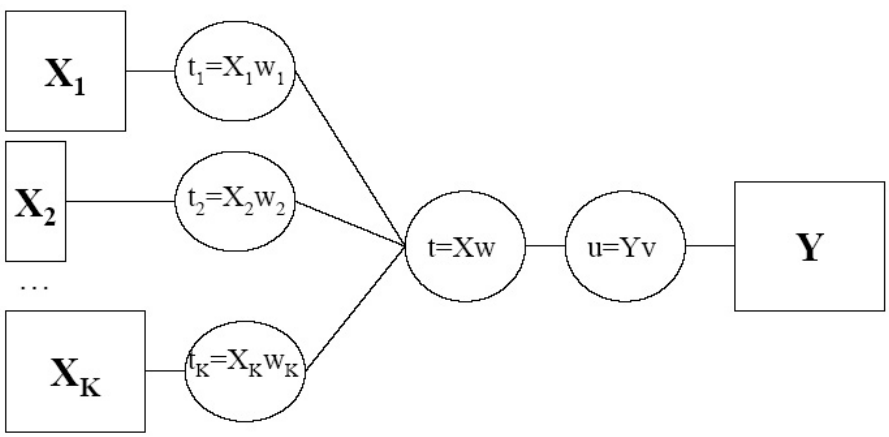
L'analyse en composantes principales sur variables instrumentales, appelée aussi analyse des redondances [Rao, 1964 ; Van Den Wollenberg, 1977 ; Sabatier, 1987], est une méthode adaptée à l'explication d'un tableau *Y* (la maladie) par un tableau *X* contenant les variables explicatives, potentiellement facteurs de risque. Des composantes, combinaisons linéaires des variables *X* orientées vers l'explication du tableau *Y*, sont utilisées pour décrire les liaisons entre les variables *X* et *Y* au travers de représentations factorielles. De plus, ces composantes peuvent être utilisées pour modéliser le tableau *Y* [Muller, 1981]. Nous proposons une méthode, appelée analyse des redondances multibloc, dérivant d'une extension de l'analyse des redondances, pour le cas où le tableau *X* est structuré en *K* sous-tableaux. Cette méthode est décrite en détail par [Bougeard *et al.*, 2007]. Tout comme l'analyse des redondances, l'analyse des redondances multibloc peut être vue sous un angle descriptif (description des liens entre variables, étude des individus), mais aussi prédictif (explication des variables *Y* par les blocs de variables explicatives).

2. VISION DESCRIPTIVE DE L'ANALYSE DES REDONDANCES MULTIBLOC

Le premier objectif de la méthode est de décrire les liens entre les variables explicatives $(X_1, \ldots, X_K)$ et les variables à expliquer *Y*. Pour cela, l'analyse des redondances multibloc recherche, dimension par dimension, une composante globale *t*, combinaison linéaire de l'ensemble des variables explicatives (*i.e.* variables du tableau concaténé $X=[X_1|\ldots|X_K]$) liée de manière optimale à une composante *u,* combinaison linéaire des variables du tableau *Y*. Cette composante globale *t* est de plus contrainte à être la synthèse des *K* composantes partielles $(t_1, \ldots, t_K)$, qui constituent elles-mêmes des résumés de chacun des tableaux $(X_1, \ldots, X_K)$. Une composante t_k contribue d'autant plus à construire la composante globale *t* qu'elle est liée à la composante *u* résumant les variables à expliquer. Les liens entre les tableaux et leurs composantes associées sont illustrés par la figure 2. Le tableau concaténé *X* joue le rôle du tableau compromis qui exprime la structure commune et spécifique des *K* tableaux X_k pour expliquer *Y*. Ce même tableau *X* fournit des composantes globales sur lesquelles se projettent l'ensemble des variables des tableaux X_k et *Y*.

Figure 2

Lien entre les K tableaux (X_1, …, X_K) et un tableau Y, résumés chacun par une composante



3. VISION PREDICTIVE DE L'ANALYSE DES REDONDANCES MULTIBLOC

Le second objectif est d'expliquer le tableau *Y* à partir de l'ensemble des variables explicatives *X*, réparties dans les *K* tableaux X_k *(k=1, …, K)*. Les composantes *t,* combinaisons linéaires de l'ensemble de ces variables *X*, sont construites de façon à être mutuellement orthogonales. Il est donc possible de les utiliser dans la régression, à la place des variables explicatives. Le retour aux variables d'origine est aisé. Le nombre de composantes à introduire dans le modèle peut être déterminé par une procédure de validation croisée [Stone, 1974]. Cette procédure consiste à diviser un grand nombre de fois (par exemple*, 500* fois) le jeu de données en deux sous-échantillons. Le premier est le jeu de données de calibration (comprenant *2/3* des individus) à partir duquel sont estimés les coefficients de régression associés au modèle liant *X* et *Y*, ainsi que l'erreur moyenne de calibration ($RMSE_C$). Le second est le jeu de données de validation (comprenant *1/3* des individus) à partir duquel est estimée l'erreur moyenne de prédiction ($RMSE_V$). Ces deux erreurs peuvent être vues comme des fonctions du nombre de composantes *t* introduites dans le modèle. Comme l'objectif de traitement des données d'épidémiologie animale est de disposer d'une méthode à la fois descriptive et prédictive, le nombre optimal de composantes est un compromis optimisant à la fois la qualité d'ajustement du modèle aux données (minimisation de l'erreur $RMSE_C$) et sa qualité prédictive (minimisation de l'erreur $RMSE_V$).

III - RESULTATS

1. DONNEES D'EPIDEMIOLOGIE ANIMALE

Les données sont issues d'une enquête analytique de type cas/témoin menée en France. L'objectif de cette étude est de déterminer les facteurs de risque du niveau de séroprévalence intra-élevage à l'égard du circovirus porcin de type 2 (*PCV2*), principal facteur infectieux impliqué dans la maladie de l'amaigrissement du porcelet, *MAP* [Rose *et al.*, 2003]. Le tableau de données comporte 158 élevages sur lesquels sont mesurées 36 variables organisées en cinq blocs, décrites par le tableau 1. Le tableau *Y*, composé de trois variables, est relatif au taux d'animaux ayant réagi à l'infection par le *PCV2* (truies, porcs âgés respectivement de 8 et 13 semaines). Les variables explicatives *X* sont organisées en quatre tableaux : X_1 relatif aux mesures de bio-sécurité et d'hygiène (10 variables), X_2 reflétant la conduite d'élevage (12 variables), X_3 lié à la structure de l'élevage (8 variables) et X_4 relatif aux co-facteurs infectieux et vaccins (3 variables). Les variables qualitatives 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, celles-ci sont centrées et réduites.

Le premier objectif de l'enquête est de décrire les liens entre les variables et entre les blocs de variables explicatives, orientés vers l'explication du niveau de séroprévalence intra-élevage. Le second objectif est à la fois de déterminer, parmi les variables de *X*, celles qui sont facteurs de risque, ou au contraire facteurs protecteurs, mais aussi de mesurer l'influence des blocs de variables (X_1, …, X_4) dans l'explication de cette même séroprévalence.

2. DESCRIPTION DE TABLEAUX STRUCTURES EN BLOCS

L'ensemble des variables peut être représenté sur un système formé par les composantes globales *t*, orthogonales mutuellement. La figure 3 illustre cette représentation sur le plan des deux premières composantes. Les variables à expliquer *Y* relatives à la proportion de truies (*CIRCOTR*) et de porcelets âgés de 8 semaines (*CIRCOPS*) séropositifs au virus *PCV2* sont liées entre elles, mais non corrélées à la proportion de porcs de 13 semaines séropositifs (*CIRCOPC*). Cette image semble proche de la réalité des faits en élevage. En effet, à 8 semaines d'âge, les porcelets disposent encore d'une immunité passive, témoin des anticorps délivrés par les truies à leur descendance, ce qui explique que leurs proportions de séroprévalence soient comparables. Les porcs à l'engrais, plus âgés, ont perdu trace de l'immunité colostrale reçue de la truie.

Tableau 1

Description des variables et des blocs de variables, du jeu de données relatif à la séropositivité d'élevages au circovirus porcin de type 2

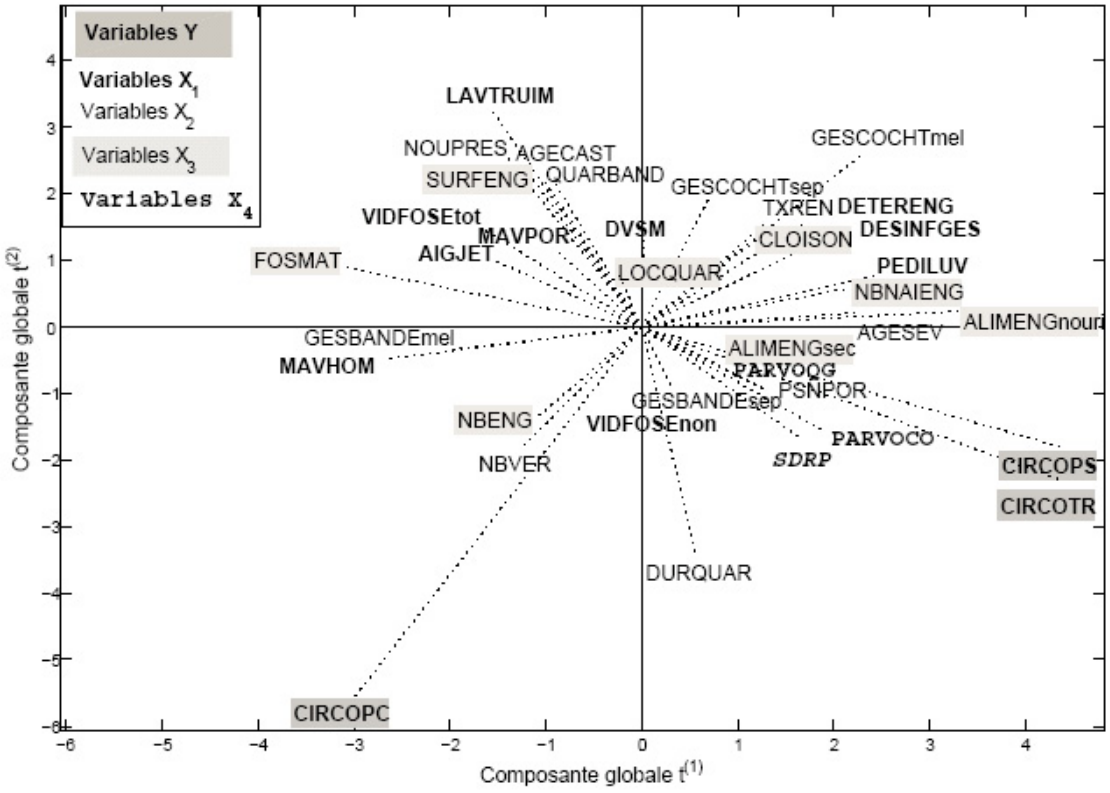
Bloc	Variable	Description
Y	CIRCOPS	Taux de porcelets positifs au circovirus *PCV2* en post-sevrage
	CIRCOPC	Taux de porcs positifs au circovirus *PCV2* en engraissement
	CIRCOTR	Taux de truies positives au circovirus *PCV2*
X_1	MAVPOR	Sens de circulation des animaux (marche en avant)
	MAVHOM	Sens de circulation pour les hommes (marche en avant)
	PEDILUV	Utilisation d'un pédiluve dans chaque salle de l'élevage
	AIGJET	Utilisation d'une aiguille jetable par truie pour les vaccins
	DETERENG	Détersion de la salle d'engraissement après lavage
	VIDFOSEnon	Vidange de la fosse (partielle *vs* pas de vidange)
	VIDFOSEtot	Vidange de la fosse (partielle *vs* totale)
	DESINFGES	Désinfection des stalles de gestation
	LAVTRUIM	Lavage des truies à l'entrée en maternité
	DVSM	Durée du vide sanitaire en maternité
X2	GESBANDEmel	Séparation physique des bandes de truies gestantes (externe *vs* mélange)
	GESBANDEsep	Séparation physique des bandes de truies gestantes (externe *vs* séparées)
	GESCOCHTmel	Position des cochettes au sein des travées (externe *vs* mélange)
	GESCOCHTsep	Position des cochettes au sein des travées (externe *vs* séparées)
	NOUPRES	Présence d'une nursery pour une partie de la bande
	AGECAST	Age des verrats à la castration
	TXREN	Taux de renouvellement du troupeau de truies
	AGESEV	Age des porcelets au sevrage
	PSNPOR	Nombre de portées par case en post-sevrage
	QUARBAND	Nombre de lots par salle en quarantaine
	DURQUAR	Durée moyenne de la quarantaine (en semaines)
	NBVER	Nombre moyen de verrats introduits par an
X3	NBENG	Nombre d'élevages engraisseurs dans un rayon de 2 km autour de l'élevage
	NBNAIENG	Nombre d'élevages naisseurs-engraisseurs dans un rayon de 2 km autour de l'élevage
	CLOISON	Cloisons entre les préfosses en engraissement
	ALIMENGnourri	Type d'alimentation en engraissement (soupe *vs* nourrisoupe)
	ALIMENGsec	Type d'alimentation en engraissement (soupe *vs* sec)
	SURFENG	Surface des cases en engraissement
	LOCQUAR	Type de locaux en quarantaine (semi-claustration ou claustration)
	FOSMAT	Profondeur des préfosses en maternité
X4	SDRP	Vaccination des truies contre le virus *SDRP*
	PARVOQG	Utilisation du même antigène contre le parvovirus en quarantaine et lors de la gestation
	PARVOCO	Taux de cochettes positives au parvovirus

Ils sont élevés dans d'autres bâtiments et sont donc soumis à un milieu différent. A ce stade, une séropositivité témoigne d'une infection active par le virus *PCV2*. Pour déterminer les facteurs de risque relatifs à la proportion d'animaux séropositifs après infection par le circovirus *PCV2*, il est essentiel de raisonner sur l'ensemble des variables *Y*. En effet, un élevage ayant un profil à risque vis-à-vis de la *MAP* est un élevage où la proportion de truies

et de porcelets séropositifs est faible (la truie transmet peu d'anticorps à ses porcelets par le colostrum) et où en revanche la proportion de porcs à l'engrais séropositifs est élevée (forte pression d'infection virale) [Rose *et al.*, 2003]. Sur la figure 3, les variables ayant des coordonnées négativement corrélées à la composante $t^{(1)}$ sont donc associées au profil d'élevage à risque. La variable *FOSMAT*, relative à la profondeur des préfosses en maternité, par exemple, est interprétée comme un facteur de risque pour l'élevage. A l'inverse, les variables ayant des coordonnées positivement corrélées à la composante $t^{(1)}$ sont associées à un profil d'élevage présentant peu de risque. La variable *PEDILUV*, définissant l'utilisation d'un pédiluve dans chaque salle de l'élevage, est interprétée comme un facteur protecteur pour l'élevage.

Figure 3

Représentation factorielle des variables sur le plan des deux premières composantes globales



3. PREDICTION A PARTIR DE TABLEAUX STRUCTURES EN BLOCS

Les résultats issus de la validation croisée montrent que quatre composantes sont nécessaires pour bien expliquer l'information contenue dans les tableaux de données, mais aussi pour garantir une bonne qualité de prédiction des variables à expliquer par les variables explicatives. L'utilisation de quatre composantes permet d'expliquer 93% de l'information contenue dans le tableau *Y* et 19% de l'information contenue dans le tableau *X*.

Une fois la dimension du modèle optimisée, il devient possible de déterminer l'influence des groupes de variables pour expliquer le niveau de séroprévalence intra-élevage au *PCV2*. Ce sont les variables relatives à la conduite d'élevage (33%) et aux mesures de bio-sécurité et d'hygiène (30%) qui influencent le plus la proportion d'animaux ayant séroconverti. Les variables liées à la structure de l'élevage (20%) et aux co-facteurs

infectieux et vaccins (17%) ont moins d'influence.

Afin de pouvoir indiquer des actions précises à mener en élevage, l'influence de chaque variable *X* dans l'explication de la séroprévalence des élevages au *PCV2* est aussi donnée. Elle est issue des coefficients de régression des variables *X* pour expliquer les variables *Y*, associés à leurs intervalles de confiance calculés par validation croisée. La détermination des facteurs de risque au niveau de l'élevage est basée sur l'interprétation conjointe des coefficients de régression associés aux trois variables *Y* à expliquer. Un élevage au profil protecteur vis-à-vis de la pression d'infection par le circovirus *PCV2* est un élevage où la proportion de porcelets et de truies présentant une séroconversion après infection par le circovirus *PCV2* est élevée, et où la proportion de porcs à l'engrais présentant une séroconversion est faible. Un facteur de risque à effet protecteur est donc une variable explicative pour laquelle le coefficient de régression est significativement positif pour les variables à expliquer *CIRCOPS* et *CIRCOTR*, et significativement négatif pour la variable à expliquer *CIRCOPC*. A l'inverse, un facteur de risque est une variable explicative pour laquelle le coefficient de régression est significativement négatif pour les variables à expliquer *CIRCOPS* et *CIRCOTR* et significativement positif pour la variable à expliquer *CIRCOPC*. L'épidémiologiste utilise l'ensemble de ces résultats pour orienter des actions à mener, en vue d'améliorer le statut sanitaire du troupeau vis-à-vis de la pression d'infection par le circovirus *PCV2*.

Afin de clarifier l'interprétation, nous donnons l'exemple du profil de l'élevage pouvant permettre d'améliorer la situation chez le porc à l'engrais (diminution de la pression d'infection par le virus *PCV2*). Du point de vue des mesures de biosécurité et d'hygiène (bloc de variables X_1), il apparaît important de laver les truies à l'entrée en maternité, pratiquer une détersion supplémentaire de la salle d'engraissement après le lavage et désinfecter les stalles de gestation. Du point de vue de la conduite d'élevage (X_2), il est préférable de positionner les cochettes à l'extérieur des travées par rapport aux truies adultes, augmenter la durée de la quarantaine, augmenter le nombre de verrats introduits dans l'élevage, augmenter le taux de renouvellement des truies, augmenter le nombre de lots par salle en quarantaine et utiliser une nursery. Du point de vue de la structure de l'élevage (X_3), le seul facteur significatif est de mettre des cloisons entre les préfosses en engraissement. Les variables relatives aux co-facteurs infectieux et vaccins n'ont pas d'influence significative.

IV - CONCLUSION ET PERSPECTIVES

Les contraintes associées au traitement statistique des données d'épidémiologie animale ont amené à développer une méthode d'analyse factorielle multibloc permettant la description de plusieurs tableaux, orientée vers l'explication d'un autre. D'autres approches ont été proposées pour l'analyse de *(K+1)* tableaux, dans des contextes privilégiant d'autres méthodes. Nous pouvons citer les travaux de [Kissita, 2003] privilégiant le cadre de l'analyse canonique, ceux de [Wold, 1984 ; Vivien, 2002] privilégiant le cadre de la régression *PLS*. Ces méthodes multiblocs sont généralement issues du domaine de la chimiométrie, mais ont aussi été appliquées sur des données de psychométrie, sensométrie, écologie, études de marché ou sciences sociales. Ce travail de recherche a permis d'appliquer les méthodes d'analyse factorielle multibloc au domaine de l'épidémiologie animale, dans lequel elles n'avaient quasiment pas été appliquées, excepté par [Rougoor *et al.*, 1999].

L'analyse des redondances multibloc proposée permet de gérer une grande partie de la complexité des données d'épidémiologie animale et des objectifs de traitements statistiques associés. La première avancée est de proposer une modélisation simultanée de plusieurs variables, ce qui permet ainsi de s'affranchir de la réalisation de plusieurs modèles aux conclusions bien souvent différentes, ainsi qu'à la synthèse des variables à expliquer en une seule variable de synthèse, solution qui n'est pas toujours satisfaisante pour l'épidémiologiste qui connaît la complexité du problème qu'il étudie. La seconde avancée apportée par ces méthodes est de permettre la prise en compte d'un plus grand nombre de variables explicatives

qu'avec les méthodes de régression usuelles. Ceci évite à l'épidémiologiste un tri délicat et peu satisfaisant. Le fait d'utiliser des méthodes factorielles couplées à des modélisations permet, de plus, d'associer les avantages de ces deux types de méthodes statistiques, trop souvent dissociés dans la pratique. La troisième et plus importante avancée est l'utilisation de méthodes peu vulnérables à la multicolinéarité. Ce point crucial est à la fois pris en compte par l'application de méthodes dont la résolution est peu sensible à la quasi-colinéarité des variables explicatives, ainsi que par l'utilisation conjointe de la régression orthogonalisée. La dernière avancée est de pouvoir désormais prendre en compte la structure en blocs des variables explicatives. Ceci permet d'apporter de nouvelles réponses à l'épidémiologiste, comme la mesure de l'influence de blocs de variables ayant un sens biologique dans l'explication de la maladie. De nombreuses possibilités de représentation graphique des résultats permettent d'enrichir l'interprétation des données.

Les perspectives de développement de ces méthodes devraient permettre de nouveaux avancements dans le traitement des données d'épidémiologie animale. Les travaux relatifs à la régression *PLS* linéaire généralisée [Marx, 1996 ; Bastien *et al.*, 2005] pourraient permettre une meilleure prise en compte des variables qualitatives. L'extension des méthodes factorielles multiblocs au cas de plusieurs tableaux à expliquer ouvrira à la possibilité d'étude des facteurs de risque de l'évolution d'une maladie au cours du temps par exemple. L'adaptation de certaines caractéristiques de méthodes, telles que l'approche *PLS* [Wold, 1982] ou *LISREL* [Joreskog, 1970], devrait permettre la prise en compte de liens entre les différents tableaux de variables explicatives. Ces perspectives, ainsi que l'augmentation du nombre de variables collectées et de la complexité des questions posées, notamment dans les domaines biologiques, devraient progressivement vulgariser l'utilisation des méthodes factorielles multiblocs.

BIBLIOGRAPHIE

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

Allison P.D. - Logistic regression using the SAS system : theory and application, Cary, NC, USA, 1999.

Barker L., Brown C. - Logistic regression when binary predictor variables are highly correlated. *Statist. Med.*, 2001, **20,** 1431-1442.

Bastien P., Vinzi V.E., Tenenhaus M. - PLS generalised linear regression. *Comput. stat. data an.*, 2005, **48,** 17-46.

Bougeard S. - Description et prédiction à partir de données structurées en plusieurs tableaux. Application en épidémiologie animale. 2007, Thèse de doctorat, Université Rennes 2, Rennes.

Bougeard S., Hanafi M., Qannari E.M. - ACPVI multibloc. Application à des données d'épidémiologie animale. *JSFdS,* 2007, **148,** 77-94.

Dohoo I.R., Ducrot C., Fourichon C., Donald A., Hurnik D. - An overview of techniques for dealing with large numbers of independent variables in epidemiologic studies. *Prev. Vet. Med.*, 1997, **29,** 221-239.

Dohoo I.R., Martin W., Stryhn H. - Veterinary epidemiologic research, Prince Edward Island, Canada, 2003.

Hosmer D.W., Lemeshow S. - Applied logistic regression, New York, 1989.

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. 2003, Thèse de doctorat, Université Paris Dauphine IX, Paris.

Lafi S.Q., Kaneene J.B. - An explanation of the use of principal-components analysis to detect and correct for multicollinearity. *Prev. Vet. Med.,* 1992, **13,** 261-275.

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

Marx B.D. - Iteratively reweighted partial least squares estimation for generalized linear

regression. *Technometrics*, 1996, **38,** 374-381.

Massy W.F. - Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 1965, **60,** 234-256.

Muller K.E. - Relationships between redundancy analysis, canonical correlation and multivariate regression. *Psychometrika*, 1981, **46,** 139-142.

Rao C.R. - The use and interpretation of principal component analysis in applied research. *Sankhya, A.*, 1964, **26,** 329-358.

Rose N., Larour G., Le Digherher G., Eveno E., Jolly J.P., Blanchard P., Oger A., Le Dinma M., Jestin A., Madec F. - Risk factors for porcine post-weaning multisystemic wasting syndrome (PMWS) in 149 french farrow-to-finish herds. *Prev. Vet. Med.*, 2003, **61,** 209-225.

Rougoor C.W., Hanekamp W.J., Dijkhuizen A.A., Nielen M.,Wilmink J.B. - Relationships between dairy cow mastitis and fertility management and farm performance. *Prev. Vet. Med.*, 1999, **39,** 247-264.

Sabatier R. - Analyse factorielle de données structurées et métriques. *Statistique et analyse des données*, 1987, **12,** 75-96.

Saporta G. - Liaisons entre plusieurs ensembles de variables et codage de données qualitatives. 1975, Thèse de doctorat, Université Paris VI.

Schaeffer R.L. - Alternative estimators in logistic regression when the data are collinear. *J. Statist. Comput. Simul.*, 1986, **25,** 75-91.

Stone M. - Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society*, 1974, **36,** 111-147.

Toma B., Dufour B., Sanaa M., Benet J.J., Ellis P., Moutou F., Louza A. - Epidémiologie appliquée à la lutte collective contre les maladies animales transmissibles majeures, Maisons Alfort, 1996.

Van Den Wollenberg A. - Redundancy analysis : an alternative for canonical correlation analysis. *Psychometrika*, 1977, **42,** 207-219.

Vivien M. - Approches PLS linéaires et non-linéaires pour la modélisation de multi-tableaux : théorie et applications. 2002, Université de Montpellier 1, Montpellier. p. 312.

Weissfeld L.A., Sereika S.M. - A multicollinearity diagnostic for generalized linear models. *Commun. Statist. - Theory Meth.*, 1991, **20,** 1183-1198.

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

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