Régression multibloc sur classes latentes. Application à l'usage d'antibiotiques en élevages de lapins

Stéphanie Bougeard, Claire Chauvin, Gilbert Saporta, Ndèye Niang

RÉGRESSION MULTIBLOC SUR CLASSES LATENTES. APPLICATION À L'USAGE D'ANTIBIOTIQUES EN ÉLEVAGES DE LAPINS.

Stéphanie Bougeard [1], Claire Chauvin [1], Gilbert Saporta [2], Ndeye Niang [2]

[1] Anses, Laboratoire de Ploufragan-Plouzané-Niort, Unité d'Epidémiologie, Santé et Bien-être, Technopôle Saint Brieuc Armor, BP53, 22440 Ploufragan, France
[2] CEDRIC-CNAM, Paris, France

RESUME *: Le traitement statistique des données d'épidémiologie analytique vétérinaire vise à déterminer les facteurs de risque d'une maladie ou d'un problème de santé publique vétérinaire. Pour répondre à cet objectif, les modèles linéaires généralisés sont utilisés. Pour le cas où les observations proviennent de différentes sous-populations, ces modèles existent sous forme de modèles sur classes latentes, aussi connus sous le nom de modèles de mélange. Cependant, en épidémiologie vétérinaire notamment, ces méthodes présentent trois principales limites : (i) le nombre d'observations dans une sous-population doit être plus grand que nombre de variables, (ii) les variables doivent présenter une distribution multinormale, et (iii) les variables ne doivent pas présenter de multi-colinéarité marquées, ces hypothèses étant rarement vérifiées en pratique. Nous proposons une extension des modèles sur classes latentes pour le cas d'un grand nombre de variables ne vérifiant pas d'hypothèse distributionnelle. Ces variables peuvent présenter, de plus, la particularité d'être organisées en blocs thématiques. La méthode proposée est appelée régression multibloc sur classes latentes. Elle combine la recherche simultanée de sous-populations au sein des observations, ainsi que de modèles de régression (multibloc) locaux associés à chacune de ces sous-populations. Cette nouvelle méthode est appliquée, à titre d'exemple, à la recherche de marqueurs de risque de l'utilisation des antibiotiques dans des élevages français de lapins.*

Mots-clés : Epidémiologie analytique, facteurs de risque, modèles de mélange, régression sur classes latentes, utilisation d'antibiotiques, lapin.

ABSTRACT*: The statistical processing of analytical epidemiological data aims to determine the risk factors for a disease or veterinary public health problem. To meet this objective, generalized linear models are used. For observations from different sub-populations, these models exist as latent class models, also known as mixture models. However, in veterinary epidemiology in particular, these methods have three main limitations: (i) the number of observations in a sub-population must be greater than the number of variables, (ii) the variables must have a multi-normal distribution, and (iii) the variables must not have strong multi-collinearity, these hypotheses being rarely satisfied in practice. We propose an extension of the mixture models for a large number of variables that do not satisfy distributional hypothesis. These variables may also have the particularity of being organized into thematic blocks. The proposed method is called multiblock regression on latent classes. It combines the simultaneous search for sub-populations within the observations, as well as local (multiblock) regression models associated with each of these sub-populations. This new method is applied, for example, to the search for risk indicators for antibiotic consumption in French rabbit farms.*

Keywords: Analytical epidemiology, risk factors, mixture models, regression on latent classes, antibiotic use, rabbit.

I. INTRODUCTION

1. CONTEXTE

Cet article traite de l'analyse statistique des données d'épidémiologie analytique vétérinaire. Cette analyse vise à déterminer les facteurs de risque liés à l'apparition et au développement d'une maladie ou d'un problème de santé publique vétérinaire. Les données recueillies par l'épidémiologiste sont constituées de variables explicatives (i.e., les facteurs de risque potentiels), souvent en grand nombre, et d'une ou plusieurs variables à expliquer (i.e., la maladie ou le problème de santé publique). Ces variables sont généralement mesurées au niveau des animaux ou des élevages qui constituent les observations. Pour déterminer les facteurs de risque de la maladie, un modèle de régression est appliqué ; il est basé sur un modèle linéaire généralisé, comme une régression logistique lorsque la variable à expliquer est décrite par deux modalités (i.e., malade / pas malade). Dans le cas général, l'épidémiologiste suppose que les observations proviennent d'une population unique et homogène et qu'aucun facteur structurel ne modifie le lien entre variable(s) à expliquer et variables explicatives. Le modèle de régression est basé sur l'ensemble des observations et les facteurs de risque sont communs à toutes les observations.

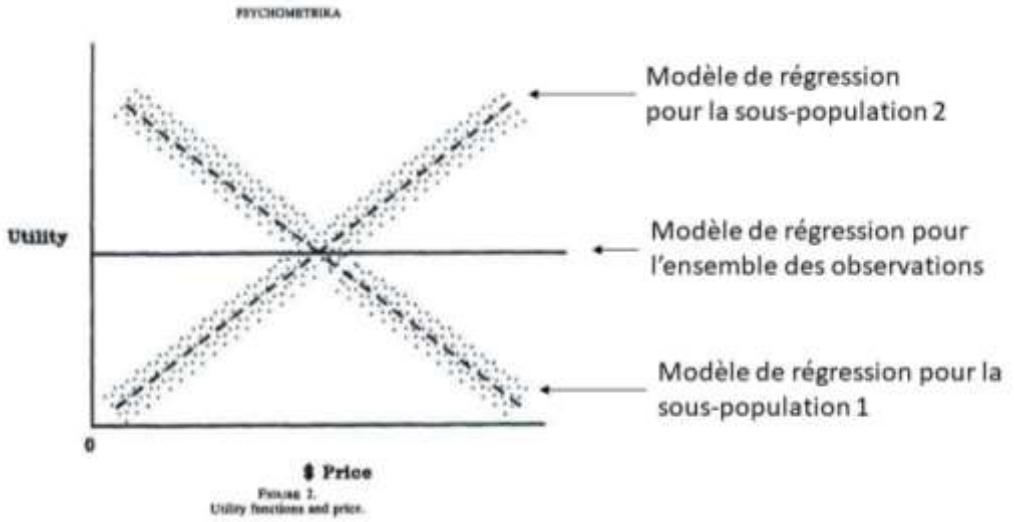
2. PROBLÉMATIQUE

Il arrive que les observations ne proviennent pas d'une population unique et homogène. Deux cas de figure existent. (i) Le premier est relatif à une hétérogénéité connue résultant de facteurs–structurant ou de confusion–qui influent sur le lien entre variable(s) à expliquer et variables explicatives. Ces facteurs sont, par exemple, les groupements de producteurs dont sont issus les élevages, ou les modes d'élevage (e.g., standard / label / extensif). Dans ce cas, il est tenu compte de ces facteurs en amont lors de la sélection (stratifiée) des observations ; ceux-ci sont ensuite intégrés au modèle de régression en tant que facteurs de confusion. Les facteurs de risque concernent l'ensemble de la population, en tenant compte des spécificités de chaque sous-population. (ii) Le second cas est relatif à une hétérogénéité inconnue de la population, tant dans le nombre de sous-populations pouvant exister, que dans la constitution de celles-ci. Ces sous-populations sont recherchées, l'objectif étant de fournir des facteurs de risque spécifiques à chacune. C'est ce second cas qui est traité dans cet article.

Dans de nombreuses applications, biologiques notamment et en épidémiologie (vétérinaire) particulièrement, il est souvent constaté un faible pouvoir explicatif du modèle de régression associé à la recherche des facteurs de risque (i.e., coefficient de détermination R^2 de valeur faible) ; ceci est généralement attribué au fait que les variables explicatives recueillies ne sont pas les plus pertinentes. Une autre cause, dont il conviendrait de tenir compte, est que les observations peuvent provenir de populations différentes et que le modèle de régression unique fournit des résultats sous-optimaux. En effet, l'estimation d'un seul ensemble de coefficients de régression pour toutes les observations peut être trompeuse car elle masque les relations réelles entre les variables, comme illustré par la Figure 1.

Figure 1

Illustration du lien entre deux variables (i.e., « price » et « utility ») : (i) pour l'ensemble des observations (ligne pleine), (ii) et pour deux sous-populations (lignes en pointillés) ; issu de DeSarbo *et al.*, 1989.



3. SOLUTIONS EXISTANTES

Une première solution consiste à réaliser une classification des observations en préalable à la régression. Un modèle de régression est ensuite construit pour chaque sous-population d'observations (e.g., Arruda *et al.*, 2016). Cette solution n'est statistiquement pas optimale car : (i) la classification est basée sur les variables explicatives (au mieux sur le tableau concaténé des variables explicatives et à expliquer), et (ii) concerne les distances mesurées dans l'espace construit par les variables, mais pas les liens impliqués dans la régression.

D'autres solutions sont apportées par les méthodes de régression sur classes latentes, aussi appelées régressions clusterwise ou typologiques. Ces modèles postulent l'existence de variables inobservables (i.e., les classes latentes ou sous-populations inconnues d'observations) dont les effets sont mesurables (i.e., les liens entre variable(s) à expliquer et variables explicatives diffèrent selon la sous-population). Ils permettent la recherche simultanée d'une partition des observations et de modèles de régression associés à ces sous-populations. Comme en régression standard, les coefficients peuvent être estimés par la méthode des moindres carrés ou du maximum de vraisemblance. La méthode des moindres carrés appliquée à la régression sur classes latentes revient à un algorithme de type K-moyennes où les sous-populations sont générées en minimisant les résidus de régression (Bock, 1969; Diday, 1976; Späth, 1979). La méthode du maximum de vraisemblance appliquée à la régression sur classes latentes suppose que les observations sont issues d'un modèle de mélange gaussien ; les méthodes qui en sont issues sont appelées modèles de mélange (DeSarbo et Cron, 1988). Elles présentent l'avantage d'être étendues aux modèles linéaires généralisés pour tenir compte du format des variables (Wedel et DeSarbo, 1995). Cependant, ces méthodes présentent trois principales limites : le nombre d'observations dans chaque sous-population doit être plus grand que le nombre de variables, les variables explicatives ne doivent pas être trop corrélées entre elles et les variables doivent présenter une distribution multinormale ; ces hypothèses sont rarement vérifiées en pratique.

Afin de lever ces contraintes, les régressions sur classes latentes ont été étendues à la régression sur composantes principales, la régression ridge (Charles, 1977), la régression PLS (Vinzi *et al*., 2005; Preda et Saporta, 2005) et la principal covariate regression (Wilderjans *et al*., 2017).

4. PROPOSITION

Nous souhaitons prendre en compte, de plus, l'organisation des variables en blocs thématiques connus *a priori*. Cette structure est usuelle en épidémiologie vétérinaire où les variables explicatives appartiennent à différents blocs que l'on souhaite différencier lors de l'analyse statistique, e.g., caractéristiques de la ferme (e.g., taille de l'élevage, performances zootechniques, autres productions animales), conduite d'élevage (e.g., taux de renouvellement, technique de reproduction, nombre d'animaux par portée), habitat (e.g., mode de ventilation, enregistrements climatiques), alimentation et abreuvement (e.g., enregistrements alimentaires, mode de distribution, nombre de mangeoires), état sanitaire du troupeau (e.g., maladies chroniques, taux de réformes, vaccinations, traitements antibiotiques), pratiques d'hygiène (e.g., protocoles de nettoyage et désinfection) et mesures de biosécurité (e.g., introduction d'animaux venant de l'extérieur, protection contre les nuisibles).

Dans ce contexte, quelques méthodes de régressions multiblocs sur classes latentes sont proposées. Parmi les plus récentes, nous pouvons citer la méthode FIMIX-PLS (Hahn *et al*., 2002), la méthode Fuzzy Clusterwise Generalized Structured Component Analysis (Hwang *et al*., 2007) et la méthode REBUS-PLS (Vinzi *et al*., 2009). Cependant, ces méthodes supposent que les blocs sont unidimensionnels (i.e., un bloc est résumé par une unique composante, ce qui revient à dire que les variables d'un bloc sont corrélées et fournissent une information consensuelle pour le cas d'une population supposée homogène) ; cette hypothèse est rarement vérifiée en épidémiologie vétérinaire.

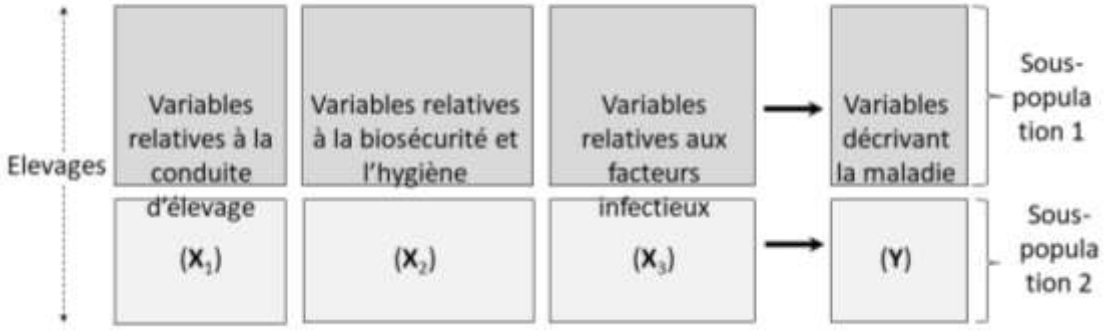
Nous proposons une extension des modèles sur classes latentes pour le cas d'un grand nombre de variables organisées en blocs thématiques, ces blocs pouvant contenir des variables résumées par plus d'une dimension, et ne vérifiant pas d'hypothèses distributionnelles. La méthode proposée combine la recherche simultanée de sous-populations inconnues au sein des observations, ainsi que les modèles de régression (multibloc) locaux associés à chacune de ces sous-populations. Cette combinaison de classification et de régression améliore la qualité du modèle de régression et facilite l'interprétation. Un test, basé sur la minimisation de l'erreur de prédiction, permet de déterminer le nombre optimal de sous-populations.

II. MÉTHODE

1. DONNÉES ET NOTATIONS

Les variables sont réparties en (K+1) blocs thématiques dont la structure est connue *a priori*, à savoir K blocs de variables explicatives $\mathbf{X} = [\mathbf{X}_1|\ldots|\mathbf{X}_K]$ et un bloc de variables à expliquer **Y**. Ces variables sont mesurées sur les mêmes N observations (e.g., les animaux ou les élevages). On suppose que ces N observations sont réparties en G sous-populations inconnues. Les blocs notés $\mathbf{X}_g$ et $\mathbf{Y}_g$ correspondent à une restriction des blocs **X** et **Y** ne contenant que les observations appartenant à la *g*ième sous-population. Cette structure est illustrée par la Figure 2.

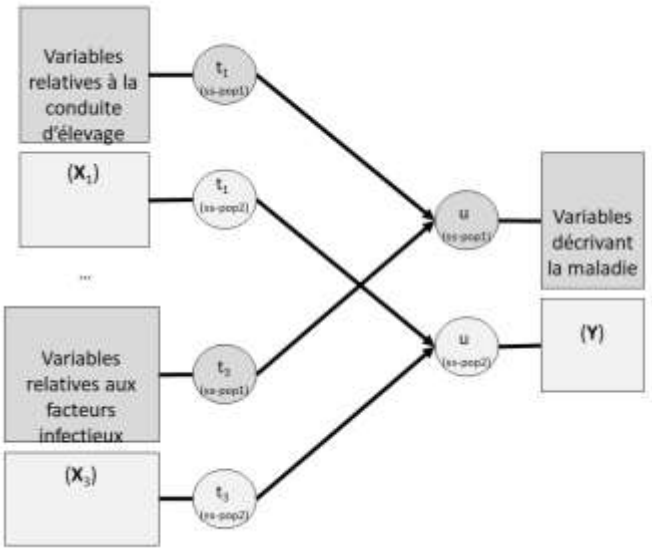
Figure 2 : Exemple de structure de données d'épidémiologie analytique vétérinaire. Les variables sont structurées en (K+1) blocs thématiques connus ; les observations sont réparties en G sous-populations inconnues. Dans cet exemple K=3 et G=2.



Afin de tenir compte du grand nombre de variables par bloc en comparaison au nombre d'observations, celles-ci sont résumées par des composantes, combinaisons linéaires des variables les constituant (généralement H=1, 2 ou 3). Ces composantes résument au mieux les variables. A titre d'exemple, si un bloc n'est composé que de variables très corrélées, une seule composante suffit à le résumer (H=1) ; si un bloc est composé de deux faisceaux de variables fournissant des informations complémentaires, deux composantes le résument (H=2). Soit $\mathbf{t}_k = \mathbf{X}_k\mathbf{w}_k$ la composante qui résume le bloc explicatif $\mathbf{X}_k$ (pour k=1, …, K) et $\mathbf{u} = \mathbf{Yv}$ la composante qui résume le bloc à expliquer **Y**. La structure des variables en (K+1) blocs, où chaque bloc est résumé par une seule composante, est illustrée par la Figure 3 pour le cas de G=2 sous-populations.

Figure 3

Exemple d'illustration du lien entre K blocs de variables (X_1, …, X_K) et un bloc Y. Chaque bloc est divisé en G=2 sous-populations d'observations. Dans cet exemple, chaque bloc –relativement à chaque sous-population– est résumé par une seule composante.



L'étape de régression est réalisée sur ces composantes (et non plus sur les variables), associée à un retour aux variables d'origine pour faciliter l'interprétation. L'étape (contraignante) de sélection des variables–pour obtenir des résultats robustes en régression standard–n'est plus nécessaire.

2. RÉGRESSION MULTIBLOC SUR CLASSES LATENTES

La méthode proposée est appelée régression multibloc sur classes latentes. Elle est basée sur la minimisation de la somme des carrés des résidus des G modèles de régression (multibloc) appliqués à chacune des G sous-populations selon le critère :

$$\min \sum_{g=1}^{G} \left\| Y_g - \widehat{Y_g} \right\|^2 \quad avec \quad \hat{Y}_g = [X_{1g,} | \dots | X_{Kg}] B_g \quad pour \quad g = (1, \dots, G)$$

Pour une sous-population g (g=1, …, G), la matrice $\widehat{Y_g}$ correspond aux valeurs estimées des variables à expliquer Y_g par le modèle de régression (multibloc). Ces valeurs sont calculées à partir des coefficients de régression B_g appliqués aux variables explicatives $X_g = [X_{1g,} | \dots | X_{Kg}]$. Les coefficients B_g sont issus d'un modèle de régression impliquant H composantes. La matrice $\left\| Y_g - \widehat{Y_g} \right\|^2$ est le carré du résidu de la régression (multibloc).

Pour trouver la partition en G sous-populations ainsi que les coefficients de régression (multibloc) $(B_1, \dots, B_g)$, l'algorithme (simplifié) de résolution proposé est le suivant :

1. Initialisation : Affectation des N observations aux G sous-populations (aléatoire ou optimisée selon une classification des données concaténées [$\mathbf{X}_1|\dots|\mathbf{X}_K|\mathbf{Y}$])
2. Pour chaque observation n :
   - Calculs de G régressions multiblocs (i.e., n appartient alternativement à chaque sous-population),
   - Pour chacun des G modèles, calcul de la somme des carrés des résidus de régression (i.e., $\sum_{g=1}^{G} \left\| Y_g - \widehat{Y_g} \right\|^2$),
   - Mise à jour de l'appartenance de l'observation n aux G sous-populations (i.e., n est affecté à la sous-population pour laquelle $\sum_{g=1}^{G} \left\| Y_g - \widehat{Y_g} \right\|^2$ est minimum)
   - Mise à jour du calcul des coefficients de régression (multibloc) locaux $(B_1, \dots, B_G)$,
3. Plusieurs initialisations sont évaluées (20 par défaut) ; la meilleure est sélectionnée (i.e., celle associée à la somme des carrés des résidus de régression $\sum_{g=1}^{G} \left\| Y_g - \widehat{Y_g} \right\|^2$ minimum).

Le détail des calculs est donné dans (Bougeard *et al*., 2018a, 2018b). De plus, il est possible de calculer, pour chaque sous-population les indices classiques fournis par les méthodes de régression multibloc, i.e., coefficients de régression, importance des blocs de variables. Afin d'en déterminer la significativité, ces indices sont associés à des intervalles de confiance calculés par simulation bootstrap. Les détails sont donnés dans (Bougeard *et al*, 2018a).

3. NOMBRE OPTIMAL DE SOUS-POPULATIONS

Il paraît intéressant que l'épidémiologiste puisse supposer et tester sans *a priori* l'existence de sous-populations inconnues dans ses observations. Ces sous-populations sont inconnues à la fois dans leur nombre et dans leur composition. L'algorithme décrit précédemment suppose que le nombre de sous-populations d'observations G et le nombre de composantes H qui résument chaque bloc de variables sont connus. Ces paramètres peuvent être choisis sur la base d'informations *a priori*, mais en général, ces valeurs ne sont pas connues. Nous proposons une procédure de sélection automatique. Pour cela, la méthode de régression multibloc sur classes latentes est appliquée–pour chaque valeur réaliste des paramètres G et H (e.g., entre 1 et 5)–au sein d'une procédure de validation croisée. Les valeurs

optimales des deux paramètres sont celles qui minimisent l'erreur de prédiction. Les détails de la procédure de validation croisée ainsi que des calculs sont détaillés dans (Bougeard *et al*., 2018b).

4. PRÉDICTION D'UNE NOUVELLE OBSERVATION

Une originalité de la méthode de régression multibloc sur classes latentes proposée, est de permettre la prédiction de la valeur d'une nouvelle observation (e.g., animal ou élevage) sur laquelle sont mesurées les mêmes variables explicatives. Cette prédiction est réalisée en deux étapes : (i) la première consiste en la prédiction de la sous-population d'appartenance à partir d'un modèle de discrimination multibloc (pour plus de détails, se référer à Bougeard *et al*, 2018b), (ii) la seconde étape consiste à prédire la(es) valeur(s) de(s) la variable(s) à expliquer à partir du modèle associé à la sous-population prédite.

III. RÉSULTATS

1. DONNÉES RELATIVES À L'USAGE D'ANTIBIOTIQUES

Les données récoltées proviennent d'une enquête rétrospective conduite en 2010 dans N=113 élevages français de lapins. L'objectif de cette enquête était d'identifier les indicateurs de risque, caractérisant les élevages, associés à leur usage d'antibiotiques. Cette étude est nécessaire dans un contexte où la rentabilité des élevages diminue et où la lutte contre l'antibiorésistance est une priorité en santé animale et humaine. La connaissance des indicateurs de risque vise à identifier des pistes d'action concrètes pour réduire l'usage des antibiotiques en élevage cunicole (Laanaya, 2011).

Les variables à expliquer (**Y**) sont les quantités d'antibiotiques utilisées sous différentes formes (injectable, dans l'eau de boisson ou dans l'aliment). Deux variables sont construites : l'utilisation d'antibiotiques sous forme de 'médicaments' d'une part (regroupant antibiotiques administrés par injection ou eau de boisson) et à travers l'aliment d'autre part. Ces deux variables, correspondent à la quantité de poids de lapin traité par antibiotique, standardisée entre les différents élevages par le poids de lapins produit au cours de la même année.

Les variables explicatives (**X**) sont issues d'un questionnaire ayant permis le recueil de 117 variables relatives aux caractéristiques structurelles de l'élevage (e.g., type de bâtiment, autres ateliers, expérience de l'éleveur), aux pratiques d'élevage (e.g., âge au sevrage, mode de rationnement) et aux performances de l'atelier cunicole. Parmi ces 117 variables, 27 ont été sélectionnées comme étant les plus liées à l'usage d'antibiotiques. Ces indicateurs de risque potentiels sont reclassés en quatre blocs thématiques relatifs aux pratiques de gestion et d'hygiène ($\mathbf{X}_1$, 8 variables), aux problèmes sanitaires ($\mathbf{X}_2$, 7 variables), à la structure de l'exploitation ($\mathbf{X}_3$, 5 variables) et aux pratiques thérapeutiques ($\mathbf{X}_4$, 7 variables). Ces variables sont décrites dans le Tableau 1.

Tableau 1

Description des variables à expliquer (Y) et explicatives (X_1, …, X_4). Enquête rétrospective relative à l'utilisation d'antibiotiques en élevage français de lapins.

Bloc	Variable	Description de la variable
Atb (Y)	Medic	Quantité d'antibiotiques utilisée sous forme de médicaments (injection ou eau de boisson)
	Alim	Quantité d'antibiotiques utilisée *via* l'aliment
Pratiques de gestion & hygiène. (X_1)	FemEng	Présence de femelles en engraissement
	FinBande	Présence d'animaux en fin de bande d'engraissement
	Nblap	Nombre de lapins produits
	VisTech	Visite du technicien d'aliment
	Chlora	Chloration permanente de l'eau en engraissement
	Ration	Rationnement par l'eau
	CtrEau	Réalisation d'autocontrôles sur l'eau de boisson
	Desinfbat	Durée du vide sanitaire après la désinfection du bâtiment
Problèmes sanitaires (X_2)	Colibacil	Présence de colibacilles sur l'élevage
	RefLapSani	Réforme des lapines pour raison sanitaire
	RefLapTec	Réforme des lapines pour raison technique
	MxPatMat	Maux de pattes en maternité
	FreqVisPb	Fréquence de visite en cas de problème
	Abces	Présence d'abcès
	MortDig	Troubles digestifs comme principale cause de mortalité en engraissement
Structure de l'élevage (X_3)	Souch1	Souche génétique des femelles 1
	Souch2	Souche génétique des femelles 2
	AgeCage	Age des cages
	Uth	Nombre d'UTH (=Unité Travail Humain) rapporté au nombre de cages
	DifProd	Différentiation de la production de lapins
Pratiques thérapeutiques (X_4)	DesinfCut	Usage de désinfectants cutanés
	Vacc	Usage de vaccins spécifiques (e.g., *E. coli*, Pasteurelles, Salmonelles)
	Vermif	Achat de vermifuges
	VisLab	Visites au laboratoire en cas de problème
	VisVet	Visites du vétérinaire en cas de problème
	Pharm	Recours à la pharmacie de l'élevage en cas de problème
	Parasito	Achat d'antiparasitaires externes

2. NOMBRE OPTIMAL DE SOUS-POPULATIONS D'ELEVAGES

Le nombre de sous-populations G, ainsi que le nombre de composantes H résumant chaque bloc, étant *a priori* inconnus, la procédure de validation croisée décrite dans le paragraphe II.3 est appliquée pour des valeurs réalistes de G et H (i.e., comprises entre 1 et 5). Les erreurs de prédiction qui en sont issues sont données dans le Tableau 2.

Tableau 2

Erreurs de prédiction selon des valeurs du nombre de sous-populations d'élevages (G=1, …, 5) et du nombre de composantes résumant chaque bloc de variables (H=1, …, 5). La valeur étoilée indique la plus petite valeur de l'erreur associée aux valeurs optimales de G et H.

Nombre de composantes (H)	Nombre de sous-populations (G)				
	G=1 ss-pop.	G=2 ss-pop.	G=3 ss-pop.	G=4 ss-pop.	G=5 ss-pop.
H=1 comp.	3,98	3,52*	3,69	3,68	3,86
H=2 comp.	3,80	3,65	3,53	3,86	3,91
H=3 comp.	3,72	3,76	3,53	3,74	3,71
H=4 comp.	3,69	3,63	3,67	3,76	3,66
H=5 comp.	3,65	3,69	4,04	4,04	3,88

La plus faible valeur de l'erreur de prédiction (=3,52) est donnée pour un modèle de régression (multibloc) appliqué à G=2 sous-populations, chaque bloc de variables étant résumé par H=1 composante. En comparaison au modèle de régression (multibloc) appliqué à l'ensemble des observations (i.e., G=1, H=1), cette erreur de prédiction est améliorée (3,98 $\rightarrow$ 3,52), ainsi que les coefficients de détermination R^2 associés. En effet, pour le modèle comportant tous les élevages, R^2=0,25 ; pour le modèle comportant deux sous-populations d'élevages, R_1^2=0,56 (sous-population 1) et R_2^2=0,65 (sous-population 2). Cette solution assigne N_1=52 élevages à la sous-population 1 et N_2=61 élevages à la sous-population 2.

3. MARQUEURS DE RISQUE DE CHAQUE SOUS-POPULATION [BLOCS]

Il est intéressant de connaître l'importance de chaque bloc de variables dans l'explication de l'usage d'antibiotiques. A titre informatif, le résultat pour l'ensemble de la population (N=113 élevages) est aussi fourni. Deux cents simulations bootstrap sont utilisées pour le calcul des intervalles de confiance ; par souci de clarté, ceux-ci ne sont pas donnés mais servent à déterminer la significativité de la valeur (i.e., significative quand la valeur 0 n'est pas comprise dans l'intervalle de confiance à 95%). Les résultats sont donnés dans le Tableau 3.

Tableau 3

Importance des blocs de variables explicatives X_k (en %) dans l'explication de l'utilisation d'antibiotiques en élevage cunicole (Y). Résultats donnés pour l'ensemble des élevages (N=113), et pour les deux sous-populations d'élevages (N_1=52, N_2=61). Une valeur étoilée indique que le bloc X_k a une importance significative dans l'explication du bloc Y ; une valeur associée au sigle « NS » indique une non-significativité de cette importance.

	Total (N=113 élevages)	Ss-population 1 (N_1=52 élevages)	Ss-population 2 (N_2=61 élevages)
Coefficient de détermination R^2	R^2=0,25	R^2=0,56	R^2=0,65
Pratiques de gestion & hyg. ($\mathbf{X}_1$, 8 var.)	26,6% *	34,6% *	31,0% *
Problèmes sanitaires ($\mathbf{X}_2$, 7 var.)	31,9% *	14,1% ^{NS}	27,4% *
Structure de l'élevage ($\mathbf{X}_3$, 5 var.)	16,5% ^{NS}	34,7% *	16,9% ^{NS}
Pratiques thérapeutiques ($\mathbf{X}_4$, 7 var.)	25,0% *	16,6% ^{NS}	24,7% *

Quelle que soit la sous-population, les pratiques de gestion et d'hygiène ($\mathbf{X}_1$) sont importantes pour expliquer l'utilisation d'antibiotiques en élevages de lapins (i.e., respectivement de 34,6% et 31,0%). Cependant, la structure de l'exploitation ($\mathbf{X}_3$; 34,7%) est importante pour les élevages de la sous-

population 1, alors que ce sont les problèmes sanitaires ($\mathbf{X}_2$; 27,4%) et les pratiques thérapeutiques ($\mathbf{X}_4$; 24,7%) qui importent pour les élevages de la sous-population 2. Par conséquent, il convient de différencier les pistes d'action, visant à réduire l'usage des antibiotiques en élevage cunicole, selon la sous-population à laquelle l'élevage appartient ; on notera qu'il est possible de prédire–selon les valeurs des variables explicatives–l'appartenance d'un nouvel élevage à l'une ou l'autre des sous-populations ainsi que son niveau d'utilisation d'antibiotique.

A titre informatif, l'interprétation réalisée sur le nombre total d'observations (N=113 élevages) fournit une interprétation comparable à celle associée à la sous-population 2, qui comporte un plus grand nombre d'élevages (N_2=61 élevages). Cependant, cette interprétation ignore les spécificités d'utilisation d'antibiotiques de la sous-population 1.

4. MARQUEURS DE RISQUE DE CHAQUE SOUS-POPULATION [VARIABLES]

Il est aussi intéressant pour l'épidémiologiste de connaître l'importance de chaque variable dans l'explication de l'utilisation d'antibiotiques, pour chacune des deux sous-populations ; à titre informatif, le résultat pour l'ensemble de la population (N=113 élevages) est donné. Deux cents simulations bootstrap sont utilisées pour le calcul des intervalles de confiance ; par souci de clarté, ceux-ci ne sont pas donnés mais servent à déterminer la significativité de la valeur (i.e., significative quand la valeur 0 n'est pas comprise dans l'intervalle de confiance à 95%). Les résultats sont donnés dans le Tableau 4.

Tableau 4

Coefficients de régression des variables explicatives dans l'explication des deux variables à expliquer, relatives à l'utilisation d'antibiotiques en élevage cunicole. Résultats donnés pour l'ensemble des élevages (N=113), et pour les deux sous-populations d'élevages (N_1=52, N_2=61). Une valeur étoilée indique que la variable explicative x a une importance significative dans l'explication de la variable y ; une valeur associée au sigle « NS » indique une non-significativité de cette importance.

Bloc	Variable	Total (N=113) Medic	Ss-pop 1 (N_1=52) Medic	Ss-pop 2 (N_2=61) Medic	Total (N=113) Alim	Ss-pop 1 (N_1=52) Alim	Ss-pop 2 (N_2=61) Alim
Coefficient de déterm. R^2		0,25	0,56	0,65	0,25	0,56	0,65
	Constante	0,00	0,22	-0,21	0,00	0,02	0,00
Pratiques de gestion & hyg, (X_1)	FemEng	-0,24 *	-0,03	-0,34 *	0,19 *	-0,03	0,44 *
	FinBande	-0,05	0,44 *	-0,21 *	0,03	0,34 *	0,28 *
	Nblap	0,31 *	0,25 *	0,34 *	-0,24 *	0,19 *	-0,44 *
	VisTech	-0,35 *	-0,11	-0,31 *	0,27 *	-0,08	0,41 *
	Chlora	0,13 *	-0,32 *	0,26 *	-0,10	-0,25 *	-0,34 *
	Ration	-0,07	0,11	-0,16 *	0,06	0,09	0,21
	CtrEau	-0,27 *	-0,48 *	-0,33 *	0,21 *	-0,37 *	0,42 *
	Desinfbat	0,25 *	0,03	0,49 *	-0,19	0,02	-0,63 *
Problèmes sanitaires (X_2)	Colibacil	0,36 *	0,29 *	0,27 *	-0,27 *	0,23 *	-0,34 *
	RefLapSani	-0,15	0,05	-0,22 *	0,11	0,04	0,28 *
	RefLapTec	-0,27 *	-0,10	-0,27 *	0,21 *	-0,08	0,35 *
	MxPatMat	-0,27 *	-0,01	-0,25 *	0,21 *	-0,01	0,33 *
	FreqVisPb	-0,15	-0,25 *	-0,08	0,11 *	-0,19 *	0,10
	Abces	-0,19 *	0,24 *	-0,44 *	0,15	0,19 *	0,57 *
	MortDig	0,41 *	-0,17 *	0,51 *	-0,31 *	-0,13	-0,66 *

Structure de l'élevage (X_3)	Souch1	0,21	0,69 *	0,23 *	-0,16 *	0,53 *	-0,29 *
	Souch2	0,14 *	-0,28 *	0,10	-0,11	-0,22 *	-0,13
	AgeCage	-0,11	0,13	-0,14	0,09	0,10	0,18
	Uth	0,25 *	-0,07	0,42 *	-0,20	-0,05	-0,54 *
	DifProd	0,36 *	-0,19	0,43 *	-0,28	-0,15	-0,56 *
Pratiques thérapeutiques (X_4)	DesinfCut	-0,32 *	0,17	-0,46 *	0,25 *	0,13 *	0,59 *
	Vacc	-0,22 *	-0,01	-0,32 *	0,17	0,00	0,41 *
	Vermif	0,19	0,20 *	-0,05	-0,14 *	0,16 *	0,06
	VisLab	0,27 *	0,13 *	0,50 *	-0,21	0,10 *	-0,65 *
	VisVet	0,16	0,31 *	0,18	-0,12 *	0,24 *	-0,23 *
	Pharm	0,20	0,33 *	0,02	-0,15 *	0,25 *	-0,02
	Parasito	0,29 *	0,07	0,23 *	-0,23 *	0,05	-0,30 *

L'interprétation des résultats de chaque sous-population est souvent différente et généralement plus précise, que celle donnée pour l'ensemble de la population. Dans cet exemple, il apparaît que : (i) dans 44,4% des cas, interpréter les liens entre variables par sous-population ou sur la population totale, n'amène pas de changement majeur (i.e., pas de modification de significativité, ni du signe des coefficients de régression). C'est par exemple le cas du lien entre la variable explicative « FemEng » et la variable à expliquer « Medic » : ce lien est significativement négatif pour l'ensemble de la population (-0,24), n'est pas significatif pour la sous-population 1 (-0,03) et est significativement négatif pour la sous-population 2 (-0,34). (ii) Dans 30% des cas, cette nouvelle interprétation fait apparaître une significativité pour au moins une sous-population, mais sans changement d'interprétation (i.e. même signe des coefficients de régression). C'est le cas du lien entre la variable explicative « RefLapSani » et la variable à expliquer « Medic » : ce lien n'est pas significatif pour l'ensemble de la population (-0,15), n'est pas significatif pour la sous-population 1 (-0,05) et est significativement négatif pour la sous-population 2 (-0,22). (iii) Dans 26% des cas, interpréter les liens entre variables par sous-population modifie l'interprétation (i.e., changement de signe des coefficients de régression). C'est par exemple le cas du lien entre la variable explicative « FinBande » et la variable à expliquer « Medic » : ce lien n'est pas significatif pour l'ensemble de la population (-0,05), est significativement positif pour la sous-population 1 (-0,44) et est significativement négatif pour la sous-population 2 (-0,21).

En n'interprétant que les variables dont les coefficients de régression sont les plus élevés (i.e., supérieur à 0,40 pour la sous-population 1 et à 0,50 pour la sous-population 2), des pistes d'action (simplifiées) pour réduire l'usage des antibiotiques en élevage cunicole peuvent être proposées. Pour la sous-population 1, limiter la présence d'animaux de fin de bande en engraissement (FinBande) et apporter une grande attention à la qualité de l'eau de boisson (CtrEau) sont avec la souche génétique (Souch1) des paramètres importants d'un moindre usage d'antibiotiques par voie injectable ou eau de boisson, tandis que la sous-population 2 est caractérisée par l'emprise des troubles digestifs (MortDig).

IV - CONCLUSION ET PERSPECTIVES

Dans cet article, nous proposons une nouvelle méthode de régression (multibloc) pour le cas d'une population dont hétérogénéité est inconnue, à la fois dans le nombre de sous-populations pouvant exister, que dans la constitution de celles-ci. La méthode proposée est appelée régression multibloc sur classes latentes. Elle combine la recherche simultanée de sous-populations au sein des observations, ainsi que de modèles de régression (multibloc) locaux associés à chacune de ces sous-

populations. Par conséquent, des facteurs de risque spécifiques à chaque sous-population sont donnés. Cette nouvelle méthode permet d'optimiser les modèles de régression (i.e., amélioration de l'explication et de la prédiction), leurs interprétations (i.e., les marqueurs de risque significatifs peuvent différer selon les sous-populations) et ainsi de conduire à de nouvelles pistes de recherche et d'action. La régression multibloc sur classes latentes proposée présente l'originalité d'autoriser la prédiction de nouvelles observations. Elle est bien adaptée au format des données d'épidémiologie vétérinaire. Un package mis à disposition sur le logiciel libre R, appelé « mbclusterwise », et comportant toutes les aides à l'utilisation, est proposé (Bougeard, 2016).

Les perspectives de développement de ces méthodes devraient permettre de nouveaux avancements dans le traitement des données d'épidémiologie vétérinaire. Plusieurs pistes de recherche peuvent être explorées. L'extension de cette problématique aux méthodes multiblocs prenant en compte des liens complexes entre blocs de variables (e.g., méthodes rGCCA, THEME, PathComDim) devraient améliorer la prise en compte de la complexité des données et des questions associées à leur analyse statistique. L'extension de la méthode aux liens autres que linéaires (e.g., logistique) devrait permettre une meilleure prise en compte des variables qualitatives, fréquentes en épidémiologie vétérinaire. Ces perspectives, ainsi que l'augmentation du nombre de variables collectées et de la complexité des questions posées, notamment dans les domaines biologiques, devrait progressivement vulgariser l'utilisation des méthodes multiblocs, sur classes latentes notamment.

BIBLIOGRAPHIE

Arruda A.G., Friendship R., Carpenter J., Hand K., Poljak Z. - Network, cluster and risk factor analyses for porcine reproductive and respiratory syndrome using data from swine sites participating in a disease control program, *Prev Vet Med*, 2016, **128**, 41-50.

Bock H.H. - The equivalence of two extremal problems and its application to the iterative classification of multivariate data, In Vortragsausarbeitung, Tagung, Mathematisches Forschungsinstitut Oberwolfach, 1969.

Bougeard, S. - Package R mbclusterwise (https://cran,r-project,org/web/packages/mbclusterwise /index,html), 2016.

Bougeard S., Abdi H., Saporta G., Niang N. - Clusterwise analysis for multiblock component methods, *Adv Data Anal Classi*, 2018a, **12**(2):285-313.

Bougeard S., Cariou V., Saporta G., Niang N. - Prediction for regularized clusterwise multiblock regression, *Appl Stoch Models Bus Ind*, 2018b, **34**(6), 852-867.

Charles C. - Régression typologique et reconnaissance des formes, *Thèse de l'Université Paris IX*, 1977.

DeSarbo W.S., Cron W.L. - A maximum likelihood methodology for clusterwise linear regression, *J Classif*, 1988, **5**, 249-282.

DeSarbo W.S., Oliver R.L., Rangaswamy A. - A simulated annealing methodology for clusterwise linear regression, *Psychometrika*, 1989, **54**(4), 707-736.

Diday, E. - Classification et sélection de paramètres sous contraintes, *Rapport technique INRIA-LABORIA*, 1976.

Hahn C., Johnson M., Hermann A.F.A. - Capturing customer heterogeneity using finite mixture PLS approach, *Schmalenbach Bus Rev*, 2002, **54**, 243-269.

Hwang H., DeSarbo W.S., Takane Y. - Fuzzy clusterwise generalized structured component analysis, Psychometrika, 2007, **72**, 181-198.

Laanaya, F. - Identification des marqueurs de risqué associés à la consommation d'antibiotiques dans les élevages de lapins, Rapport de stage de Master 2, Université de Rennes 2, 2011.

Preda C., Saporta G. - Clusterwise PLS regression on a stochastic process, *Comput Stat Data Anal*, 2005, **49**, 99-108.

Späth H. - Clusterwise linear regression, *Computing*, 1979, **22**, 367-373.

Vinzi V., Lauro C., Amato S. - PLS typological regression, In: *New developments in classification and data analysis*, Vichi M., Monari P., Mignani S., Montanari A. (Eds), Springer, 2005, 133-140.

Vinzi V., Trinchera L., Squillacciotti S., Tenenhaus M. - REBUS-PLS: a response-based procedure for detecting unit segments in PLS path modeling, *Appl Stochastic Models Bus Ind*, 2009, **24**, 439-458.

Wedel, Michel; DeSarbo, Wayne S. (1995). "A mixture likelihood approach for generalized linear models." Journal of Classification 12(1): 21-55.

Wilderjans T.F., Vande Gaer E., Kiers H.A.L., Van Mechelen I., Ceulemans E. - Principal covariates clusterwise regression (PCCR): accounting for multicollinearity and population heterogeneity in hierarchically organized data, *Psychometrika*, 2017, **82**(1), 86-111.