

ÉTUDE DES LÉSIONS PULMONAIRES CHEZ LE PORC : UNE ANALYSE SYMBOLIQUE SUR DES CONCEPTS ISSUS DE L'APPROCHE CLASSIQUE

Christelle FABLET¹, Carole TOQUE², Stéphanie BOUGEARD¹, Edwin DIDAY³

¹Afssa-Site de Ploufragan, Unité d'Epidémiologie et de Bien-Etre du Porc, Zoopôle Beaucemaine, 22440 Ploufragan France c.fablet@AFSSA.FR, s.bougeard@AFSSA.FR

²Syrokko, Aéroport de Roissy, Bat. Aéronef, 5 rue de Copenhague, 95731 Roissy Charles de Gaulle Cedex toque@syrokko.com

³CEREMADE Université Paris Dauphine 75775 Paris Cedex 16 France edwin.diday@ceremade.dauphine.fr

Résumé. L'objectif de l'étude est de décrire les liens qui peuvent exister entre des variables quantitatives et qualitatives caractérisant les lésions pulmonaires chez les porcs, et plus particulièrement les deux maladies que sont la pneumonie et la pleurésie. Les données présentent une structure emboîtée et consistent en 125 élevages comprenant chacun 30 porcs. Une approche classique, nécessitant la création de nombreuses variables agrégées telles que des moyennes et des médianes au niveau des élevages, basée sur l'analyse en composantes principales (ACP) suivie d'une classification ascendante hiérarchique (CAH), conduit à la formation de quatre classes d'élevage. Parallèlement, une analyse de données symboliques dont une ACP, a été conduite principalement sur des variables à valeur histogramme correspondant directement aux fréquences des porcs de chaque élevage. Nous concluons à la pertinence de l'analyse de données symboliques appliquée grâce au logiciel SYR en épidémiologie animale. En particulier, l'analyse de données symboliques permet de prendre en compte la variabilité des données initiales, et les différences entre classes sont identifiées grâce au logiciel qui permet de trier les élevages et les variables sur la base des fréquences de leurs modalités en s'affranchissant de la création de nombreuses variables de synthèse.

Abstract. The objective is to relate quantitative and qualitative variables describing lung lesions in pigs, especially pneumonia and pleuritis. The data consist of 30 pigs from each of 125 farms. A classical approach involving several aggregated variables at the farm level, based on principal component analyses (PCA) and a hierarchical ascendant classification (HAC), lead to four classes of farms. Then, a symbolic data analysis (SDA) with a PCA was conducted on the resultant histograms.

We conclude on the relevance of SDA and the SYR software to veterinary epidemiology. In particular, SDA is able to understand the variability of initial data, and differences between classes are identified by the software's capability to order farms and variables based on the frequencies of merged categories.

Mots-clés. Epidémiologie, données emboîtées, variables histogrammes, logiciel SYR, analyse de données symboliques.

Introduction

L'épidémiologie animale vise à étudier les maladies et les facteurs de santé dans une population animale (Toma et al., (1996)). Les maladies sont souvent complexes et décrites par plusieurs variables à expliquer (taux de mortalité, signes cliniques, lésions observées en élevage et à l'abattoir). Dans la population animale considérée, les informations relatives

aux variables à expliquer peuvent être collectées à l'échelle de l'individu (animal), du groupe d'animaux de même âge (bande), et/ou de l'élevage. Dans les premières étapes d'étude d'une maladie, deux objectifs d'enquêtes épidémiologiques sont de (i) décrire les liens entre les variables caractérisant la maladie et (ii) classer l'échantillon de population étudiée en différents gradients de sévérité d'atteinte vis-à-vis de la maladie. Sur le plan méthodologique, une analyse de données descriptive est souvent effectuée. Celle-ci se fonde, selon la nature des données à traiter, sur une Analyse Factorielle des Correspondances Multiples (AFCM) ou une Analyse en Composantes Principales (ACP) suivie d'une méthode de classification telle que la Classification Ascendante Hiérarchique (CAH). Nous considérerons ce type d'analyse de données comme une *approche classique*. Lorsque l'unité d'étude est le groupe d'animaux, i.e. la bande, une partie ou la totalité des variables décrivant la maladie est enregistrée sur un échantillon d'animaux de la bande et les valeurs varient selon les individus. Des méthodes prenant en compte la structure hiérarchique des données, comme par exemple les modèles de mélanges linéaires en régression logistique, ont été utilisées en épidémiologie animale (Dohoo et al., (2003)). Toutefois, dans le domaine vétérinaire, celles-ci sont exclusivement mises en œuvre pour mesurer l'intensité de la liaison entre des variables explicatives et la maladie à expliquer. A notre connaissance, aucune de ces méthodes ne vise à effectuer un classement des individus.

L'analyse des données par l'*approche classique* utilise des moyennes de valeurs de l'échantillon, la variabilité des données initiales étant ainsi perdue. D'une part, la création de ces nombreuses variables suivie du tri des plus pertinentes alourdit nettement l'analyse, et d'autre part, la perte d'information conduit forcément à des imprécisions ainsi qu'à des erreurs de classement et d'interprétation. En effet, une valeur moyenne identique entre deux groupes d'individus peut correspondre en réalité à des distributions de valeurs différentes, point essentiel à considérer dans la compréhension des phénomènes de santé. Ces dernières années, une méthode d'analyse de données symboliques accompagnée d'un logiciel a été développée (Diday et Noirhomme (2008), Billard et Diday (2006)). Les données sont dites *symboliques* parce qu'elles conservent la description de la réalité dans toutes ses variations, sous forme d'histogrammes, d'intervalles ou de courbes de répartition. Cette approche permet notamment de définir des variables plus riches sur des concepts ou classes obtenues à l'issue de procédures d'agrégation, et a trouvé des applications dans plusieurs domaines tels que la santé humaine, le bâtiment, la finance. Son intérêt en épidémiologie animale reste donc à évaluer. Pour ce faire, un jeu de données relatif aux lésions pulmonaires chez les porcs en croissance a été utilisé. Les données ont été collectées dans le cadre d'une enquête épidémiologique analytique qui a pour finalité d'identifier les facteurs de risque de deux maladies affectant le poumon : la pneumonie et la pleurésie.

L'objectif du présent travail est d'optimiser la description des liens qui peuvent exister entre des variables quantitatives et qualitatives décrivant les lésions pulmonaires (i) en commençant par une partition basée sur l'approche classique de type analyse en composantes principales suivie d'une classification ascendante hiérarchique et (ii) en améliorant les résultats de l'analyse classique par une approche symbolique qui utilise des variables plus riches sur des classes ou concepts.

1 Données d'épidémiologie animale

Le jeu de données comporte 125 élevages. Pour chacun d'entre eux, des mesures sont réalisées sur 30 porcs par l'observation de leurs poumons à l'abattoir. Sur ces poumons, différentes notations de lésions sont effectuées : *Pneumonie* : note qui varie de 0 (absence de lésion) à 28 (affecte tout le poumon) pour le porc, issue d'une notation (0 à 4) de chacun des 7 lobes pulmonaires (Madec et Kobisch (1982)) ; *Pleurésie* : note de 0 (absence de lésion) à 4 (pleurésie affectant l'intégralité du poumon) (Madec and Kobisch (1982)) ; *Abcès pulmonaires* : variable qualitative dichotomique : (présence, absence) ; *Nodules pulmonaires* (induration) : variable qualitative dichotomique (présence, absence) ; *Lésions*

cicatricielles de pneumonie : variable qualitative dichotomique (présence, absence) ; *Hypertrophie et congestion des ganglions lymphatiques pulmonaires* : note de 0 (absence) à 3 (hypertrophie ou congestion importante) ; *Péricardite* (inflammation du péricarde) : variable qualitative dichotomique (présence, absence). Puis, chaque élevage est décrit par 2 autres variables relatives à l'intensité des signes cliniques respiratoires (fréquence de toux) dans deux bandes de porcs.

2 Approche classique

2.1 Données calculées et ACP successives

Afin d'étudier les relations entre les différentes variables décrivant les maladies, plusieurs variables ont été créées au niveau de l'élevage à partir des données individuelles (porc). Ces variables sont la note moyenne, la note médiane de pneumonie du lot (élevage) ainsi que les fréquences de porcs présentant chaque note de pneumonie (ex : % de porcs avec note à 0, 1 ..., % de porcs avec note comprise entre 1 et 4) et les fréquences de porcs présentant les différentes lésions. Au total, chaque élevage est décrit par 64 variables. Afin de déterminer la combinaison de variables permettant de discriminer au mieux les élevages à l'égard de la pathologie pulmonaire, plusieurs analyses en composantes principales ont été effectuées successivement. L'objectif est de réduire le nombre de variables pour retenir les variables contribuant le plus à la formation des axes et celles correctement représentées (\cos^2). Au final, 17 variables ont été retenues.

2.2 Groupes de variables

Les résultats du cercle des corrélations des 17 variables sur le plan 1-2 (fournissant 50.6% de l'inertie) sont présentés en Figure 1.

Les variables décrivant un faible (respectivement haut) niveau d'atteinte à l'égard des 2 maladies affectant le poumon sont localisées sur la partie gauche (respectivement droite) du cercle des corrélations. La partie droite du cercle montre deux groupes de variables caractérisant, en haut à droite la sévérité des lésions de pneumonie (PNEU +) et les descripteurs cliniques associés, et en bas à droite les variables caractérisant principalement la pleurésie (PLEU +). La partie gauche du cercle montre un groupe de variables décrivant un faible niveau d'atteinte pour les deux maladies pulmonaires (Pulm-).

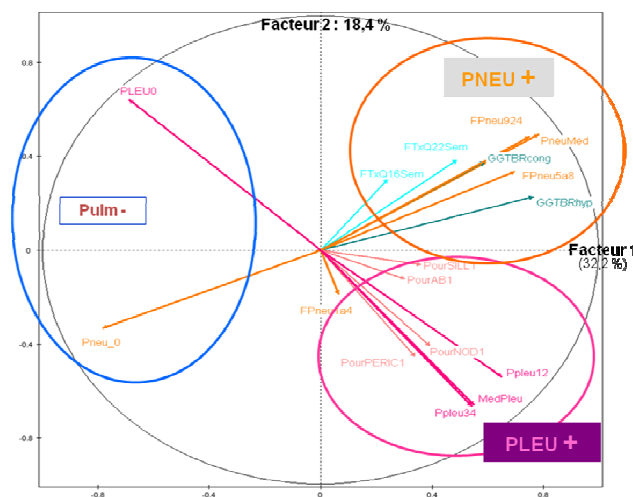


Figure 1: Analyse classique : Cercle des corrélations pour les 17 variables, et les groupes de variables

2.3 CAH et description classique des classes

En complément des résultats de l'ACP, une Classification Ascendante Hiérarchique (CAH) a été réalisée afin de former des groupes ou classes d'élevages homogènes. La classification est effectuée à partir des 17 variables utilisées dans la dernière ACP. Au final, 4 groupes d'élevages (comprenant respectivement 31, 12, 53, et 29 élevages) sont

formés. Les classes 1 et 3 sont formées d'élevages dont les porcs souffrent principalement de pneumonie. Cependant, les élevages de la classe 1 présentent une forme plus sévère de pneumonie en terme de fréquence de porcs affectés et d'étendue des lésions (notes ≥ 5) que ceux formant la classe 3. La classe 2 s'oppose aux autres classes en terme de type de maladie. Les élevages sont principalement caractérisés par une atteinte des séreuses et plus particulièrement par de la pleurésie. Enfin, la classe 4 correspond aux élevages les moins affectés par des troubles respiratoires.

3 Analyse symbolique sur les élevages (ou concepts)

Les descriptions des variables symboliques ainsi que de leurs méthodes d'analyse, sont détaillées dans Billard et Diday (2006), Diday et Noirhomme (2008), et Diday (2010).

3.1 Description symbolique

Pour les variables quantitatives et qualitatives, des histogrammes associés à chaque élevage ont été générés à l'aide du module TABSYR du logiciel SYR. Ils correspondent aux fréquences des porcs de l'élevage. Ainsi, chaque élevage, comportant au maximum 30 individus (porcs), a été décrit au travers de 19 variables symboliques et une variable supplémentaire (Note de pneumonie en 8 classes).

Comme première approche pour décrire ces variables, nous avons procédé à une ACP des modalités des 20 variables à valeur histogramme. A cet effet, le module STATSYR du logiciel SYR a été utilisé.

3.2 ACP de variables à valeur histogramme

Nous présentons en Figure 2 les résultats du cercle des corrélations pour 8 d'entre elles, avec leurs modalités les plus représentatives. Il apparaît 4 groupes de variables dont 2 axes majeurs : l'axe PNEU+/PNEU- et l'axe PLEU_PNEU/PNEU0_PLEU0. Comparativement au groupe Pulm- issu de l'analyse classique (Figure 1), le groupe de variables PNEU0_PLEU0 est mieux défini par des variables plus corrélées entre elles pour laisser apparaître un groupe distinct PNEU-. Par ailleurs, le groupe de variables PLEU_PNEU, caractérisé par des variables à moyenne et forte pleurésie, et apparaissant sous l'étiquette PLEU+ lors de l'analyse classique (Figure 1), s'est ici enrichi de variables caractérisant une pneumonie moyenne avec, par exemple, les modalités 2 et 3 de la variable symbolique NotePneu8. A la lecture de ce seul cercle des corrélations, une description plus détaillée des groupes de variables grâce à leurs modalités respectives est donnée.

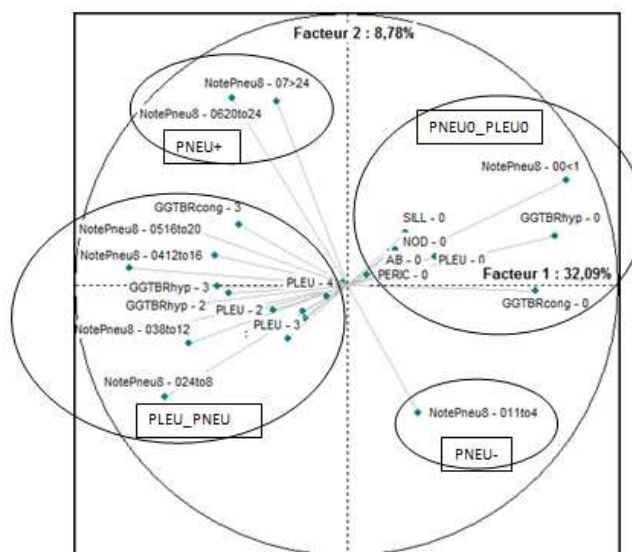


Figure 2: *Analyse symbolique* : Cercle des corrélations pour huit variables à valeur histogramme, et les groupes de variables

3.3 Analyse symbolique des classes issues de la CAH

Les quatre classes formées par la CAH ont été analysées par une approche symbolique afin d'optimiser la description des classes en tenant compte de la variabilité des données. Les numéros d'attribution des élevages aux quatre classes formées par la CAH ont été récupérés. Une variable « Numéro de la classe » est créée et constitue l'identifiant des élevages. Puis, une analyse symbolique descriptive est effectuée sur les 20 variables. Les résultats de l'analyse peuvent être donnés sous forme numérique ou graphique comme en Figure 3 (module TABSYR de SYR). Les variables y sont ordonnées en fonction de leur pouvoir discriminant vis-à-vis des quatre classes. Au sein des 20 variables, la note de pneumonie totale (NotePneu28, NotePneu8), qu'elle soit regroupée en 8 classes ou non, apparaît comme la variable la plus discriminante des groupes. Les variables relatives à la clinique respiratoire ou celles concernant les notes de pneumonie des lobes cardiaques droits et gauches, et apical droit sont également importantes pour distinguer les classes. Par exemple, concernant les variables décrivant la sévérité d'atteinte à l'égard de la pneumonie, la classe 4 (Pneu-/Pleu-) est caractérisée par les plus faibles fréquences de porcs atteints de notes élevées, et la classe 1 (Pneu +++) par les fréquences les plus importantes. Puis, une augmentation progressive de la fréquence de porcs atteints de lésions étendues de pneumonie est notée de la classe 4 à la classe 3, puis 2, puis 1.

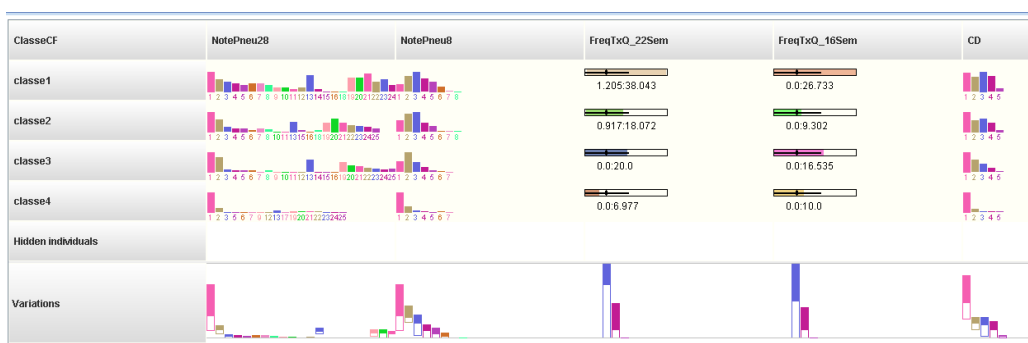


Figure 3: Description symbolique des quatre classes issues de la CAH. Les cinq premières variables sont ordonnées selon leur pouvoir discriminant par ordre décroissant.

Le logiciel SYR (module STATSYR) permet de caractériser les classes formées par les valeurs des variables les plus et les moins discriminantes de la classe comme illustré pour la classe 2 (Pleu) en Figure 4. Une identification des classes opposées pour les modalités des variables les plus discriminantes peut être aussi établie. Ainsi, la classe 2 est caractérisée par des notes élevées de pneumonie et de pleurésie. Les notes 21 et 24 sont 4 fois plus fréquentes dans cette classe. Sur ces modalités, la classe 2 s'oppose à la classe 1. La note de pleurésie de 3, la présence d'abcès, de nodules pulmonaires et de péricardite sont environ 3 fois plus fréquentes dans la classe 2. Ceci l'oppose aux élevages de la classe 4, considérée comme la plus faiblement concernée par des lésions pulmonaires.

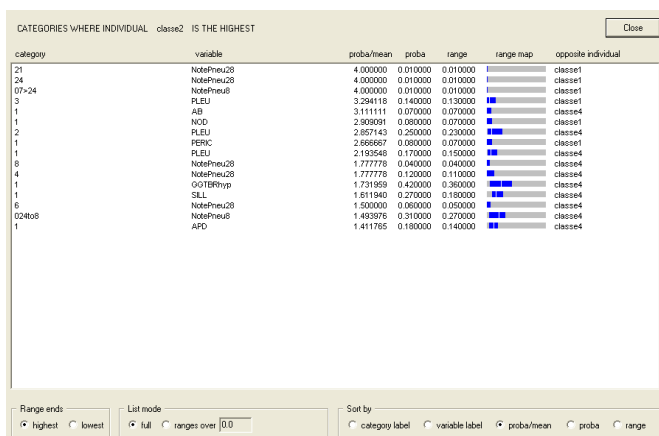


Figure 4: Une caractérisation de la classe 2 (Pleurésie avec Pneumonie) selon les modalités des variables les plus discriminantes.

3.4 Comparaison des résultats et signification biologique

Au regard de ces résultats, l'analyse symbolique des classes obtenues par la méthode classique, conforte et enrichit les conclusions établies au terme de l'analyse classique (ex : classe 2). En effet, le gradient de sévérité d'atteinte des élevages à l'égard de la pneumonie, allant de la classe 4 (la moins affectée par des troubles pulmonaires) en passant par la classe 3 (modérément atteinte) à la classe 1 (la plus sévèrement touchée) établi lors de l'interprétation de l'analyse classique, est retrouvé dans l'approche symbolique. Les résultats symboliques confirment la position particulière de la classe 2 à l'égard de la pleurésie avec de plus un complément d'information concernant l'apparition de niveaux de sévérité d'atteinte en pneumonie (notes 21 et 24 sont 4 fois plus fréquentes dans cette classe par rapport à la moyenne). De plus, ces résultats permettent de positionner ces élevages à l'égard de la pneumonie, point qui n'était pas précisé dans l'analyse classique. En effet, l'analyse symbolique des classes permet de caractériser celles-ci par l'ensemble des variables descriptives tandis que les informations obtenues par la classification concernent essentiellement les variables les plus discriminantes de chaque classe. Certaines variables ou certaines modalités n'apparaissent pas comme essentielles dans l'approche classique. La description symbolique des classes permet de compléter et préciser les informations de l'analyse classique. Elle présente le double avantage de quantifier l'importance des modalités caractéristiques et de déterminer les oppositions inter classes comme le logiciel SYR le permet en triant les élevages et variables tout en respectant les fréquences des modalités. Un avantage essentiel de l'analyse de données symbolique en comparaison à l'approche classique est une approche directe sur les données brutes. En effet, cette analyse évite de créer un grand nombre de variables qu'il faut sélectionner par la suite, ainsi que des pertes d'informations et imprécisions dues à la synthèse des données de base.

Conclusion

En conclusion, l'analyse des données symboliques semble particulièrement adaptée au traitement des données d'épidémiologie animale, grâce à l'ACP sur la base d'un tableau d'histogrammes, ainsi que dans les phases de description des classes ou concepts. En comparaison à l'analyse classique, son double avantage réside dans la prise en compte de la variabilité des données initiales et dans sa capacité à quantifier l'importance des variables dans la classification. Elle permet de traiter des tableaux de données comportant des variables à différents niveaux d'agrégation (individuel et groupe), sans réduire l'information, pour les variables collectées au niveau individuel, à un paramètre de distribution (moyenne ou médiane).

Bibliographie

- [1] Billard, L. et Diday, E. (2006) *Symbolic Data Analysis: Conceptual statistics and data Mining*, Wiley series in computational statistics, London.
- [2] Diday, E. et Noirhomme-Fraiture M. (2008) *Symbolic Data Analysis and the SODAS software*, Wiley series in computational statistics; London.
- [3] Diday, E. (2010) *Principal Component Analysis for categorical histogram data: Some open directions of research*, Springer - Serie Classification and Data Analysis, Proceedings SFC-CLADAG 2008.
- [4] Dohoo, I.R., Martin, W. et Stryhn, H. (2003) *Veterinary epidemiologic research*, University of Prince Edward Island, Prince Edward Island, Canada.
- [5] Madec, F. et Kobisch, M. (1982) *Bilan lésionnel des poumons de porcs charcutiers à l'abattoir*, Journées de la Recherche Porcine, 14, 405-412.
- [6] Toma, B., Dufour, B., Sanaa, M., Benet, J.J., Ellis, P., Moutou, F. et Louza, A. (1996) *Epidémiologie appliquée à la lutte collective contre les maladies animales transmissibles majeures*, Association pour l'Etude de l'Epidémiologie des Maladies Animales, Maisons-Alfort.