LA TYPOLOGIE, UN OUTIL AU SERVICE DU DIALOGUE ENTRE ÉPIDÉMIOLOGISTE ET PRATICIEN. L'EXEMPLE DE L'ANALYSE DES PROFILS SÉROLOGIQUES COMPLEXES *

Vincent Auvigne[1] et **Lucas Léger**[1]

RÉSUMÉ

La typologie est une technique statistique d'analyse de données qui vise à regrouper les individus étudiés en types, ou catégories, en fonction de leur ressemblance sur un ensemble de variables. C'est un outil adapté à l'étude de réalités complexes. L'objet de cette communication est de présenter une application de cette technique à un cas d'analyse de résultats de profils sérologiques, constitués de cinq analyses spécifiques de sérotype d'un agent infectieux (ELISA vis-à-vis de cinq antigènes d'*E. coli* chez le porc). La procédure d'analyses de données consiste en une analyse en composantes principales suivie d'une classification hiérarchique ascendante. Les groupes sont décrits et la signification biologique de la typologie discutée.

L'intérêt de cette approche est discuté. La typologie peut en particulier être un outil intéressant au service du dialogue entre épidémiologiste et praticien car elle combine rigueur statistique et approche centrée sur l'individu. Elle peut également être la base d'outils d'aide au diagnostic.

Mots-clés : méthodologie, typologie, sérologie.

SUMMARY

Cluster analysis is a statistical method designed to group individuals in clusters based on their similarity with respect to a set of variables. This method is well fitted to analyze complex situations. This paper presents an application of this method to the interpretation of seroprofiles in some sow herds. Each seroprofile was defined based on the results of five serotype-specific Elisa tests against five *E. coli* antigens. The analysis included a Principal Components Analysis followed by a Hierarchical Clustering. Clusters are described and the biological interpretation of the classification obtained is discussed.

Cluster analysis is a useful tool for a more productive dialog between epidemiologists and clinicians to the extent that it is statistically based and, at the same time, focused on individuals. It can be used to build diagnostic support tools.

Keywords: Methodology, Typology, Serology.

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

[1] Ekipaj, 22 rue d'Assas, 49000 Angers, France

I - INTRODUCTION

Le clinicien et l'épidémiologiste sont deux acteurs intervenant dans la gestion de la santé animale mais leurs approches diffèrent et ils ont parfois du mal à se comprendre. Une de ces différences serait un rapport différent à la généralisation et à l'individualisation : l'approche épidémiologique est centrée sur la population et l'approche clinique sur l'individu (l'individu pouvant être un troupeau) [Auvigne et Belloc, 2012]. Dans ce cadre, l'objet de cette communication est de présenter une méthode d'analyse de données, l'analyse typologique, et de discuter de son apport à l'intégration des visions de l'épidémiologiste et du clinicien.

Un cas pratique est utilisé pour cette présentation : celui du diagnostic de la qualité de la vaccination contre les diarrhées à *Escherichia coli* chez le porc. Cette vaccination est pratiquée chez les truies avec l'objectif de protéger les porcelets par voie passive *via* le colostrum. La qualité de la vaccination peut être évaluée en titrant les anticorps présents dans le colostrum. Elle a été développée par MSD santé animale et consiste en la réalisation de six différentes Elisa spécifiques [Riising *et al.*, 2005]. Ces Elisa sont spécifiques des anticorps dirigés contre les cinq antigènes fibrillaires et l'entérotoxine présents dans le vaccin produit par ce laboratoire (PORCILIS® AR-T DF et NOBI®-VAC AR-T). Pour évaluer la qualité de la vaccination dans un élevage, le protocole consiste à prélever et analyser le colostrum de 15 truies de différents rangs de portée. La moyenne arithmétique est ensuite calculée pour chacune des six Elisa. Il s'avère cependant que l'interprétation des résultats de ces profils sérologiques est délicate, voire déroutante, car l'interprétation séparée des titres des six Elisa d'un même élevage aboutit régulièrement à des conclusions discordantes. Une approche fondée sur l'analyse typologique a donc été développée pour résoudre cette difficulté.

II - MÉTHODOLOGIE DE RÉALISATION DE L'ANALYSE TYPOLOGIQUE

1. DONNÉES DISPONIBLES

L'analyse est réalisée à partir des résultats de 76 profils sérologiques (1 263 colostrums au total) réalisées dans des élevages français entre 2006 et 2012. Le protocole de vaccination dans les élevages est connu, mais l'information est parfois incomplète, en particulier quand des changements de protocoles ont été effectués. A cette base de données ont été ajoutés les résultats de quatre lots suivis dans des conditions expérimentales et dont le protocole vaccinal est parfaitement connu (deux lots vaccinés et deux lots témoins) [Besson *et al.*, 2010 ; Riising *et al.*, 2005].

L'analyse des corrélations entre les titres montre que les différents titres sont tous corrélés entre eux (tableau 1). Ces corrélations ne sont cependant pas parfaites, ce qui est cohérent avec l'existence de « discordances » signalées par les praticiens interprétant les profils.

2. CONSTRUCTION DU MODÈLE

L'objectif de l'analyse typologique est de créer des groupes d'individus (les élevages dans le cas présent) de façon à ce que les individus soient semblables au sein des groupes et différentes de celles des autres groupes. Aucune hypothèse n'est posée sur le nombre de groupes ou sur leurs caractéristiques. L'ensemble de ces analyses ont été effectuées avec le logiciel R (version 2.13.1, bibliothèque ade4, fonctions dudi.pca, dist, hclust, cutree, kmeans).

Tableau 1

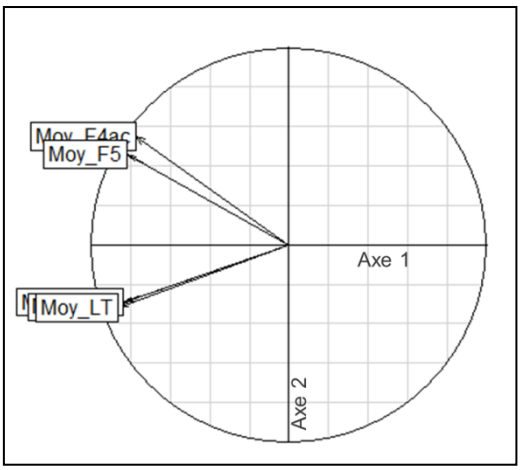
Corrélation entre les titres *E. coli* (moyenne arithmétique par élevage)

	F4ab	F4ac	F5	F6	LT
F4ab	1	0.58	0.46	0.79	0.68
F4ac	0.58	1	0.78	0.54	0.41
F5	0.46	0.78	1	0.57	0.65
F6	0.79	0.54	0.57	1	0.84
LT	0.68	0.41	0.65	0.84	1

La première étape consiste en la réalisation d'une analyse multidimensionnelle. Dans le cas d'étude, il s'est agi d'une analyse en composantes principales (ACP, Figure 1) car les variables à étudier sont quantitatives. L'ACP permet d'étudier la variabilité globale du jeu de données et de créer de nouvelles variables qui ne sont pas corrélées entre elles. Ces nouvelles variables sont tout simplement les axes de l'ACP, et les valeurs prises par ces variables sont les coordonnées des individus (ici les élevages) sur les différents axes. L'étude du pourcentage de variabilité expliqué par chaque axe (chaque nouvelle variable) et le sens biologique que l'on peut y lire permettent de déterminer combien d'axes (de variables) seront retenus pour la suite de l'analyse. Il n'existe pas de règle stricte pour effectuer ce choix, il dépend en grande partie de l'expertise de l'analyste. Dans le cas étudié, les trois premiers axes ont été retenus, ils expliquent 96 % de la variabilité du jeu de données. Le premier axe, qui explique à lui seul 71 % de la variabilité, oppose les élevages selon leur réponse globale aux six Elisa (globalement forte ou globalement faible). Le deuxième axe synthétise une opposition entre les réponses aux Elisa F4ac et F5, d'une part, et F4ab, F6 et LT, d'autre part. Le troisième axe synthétise une opposition entre les réponses aux Elisa LT et F5, d'une part, et F4ab et F4ac, d'autre part.

Figure 1

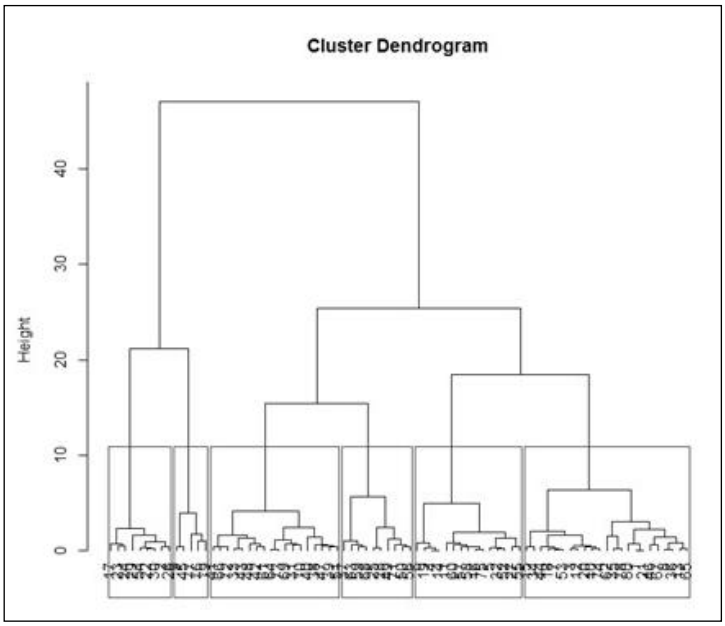
Première étape : analyse en composantes principales (ACP)



La deuxième étape consiste en une classification ascendante hiérarchique (CAH, Figure 2). L'objectif de la CAH est de constituer les groupes. Elle est fondée sur le calcul de distances (ici euclidiennes) entre les individus puis à leur assemblage en groupes successifs, en fonction de leur proximité, jusqu'à ce que tous les individus soient rassemblés dans un seul groupe. Les variables utilisées pour réaliser cette CAH sont les axes de l'ACP qui ont été retenus à l'étape précédente. A l'issue de la CAH un dendrogramme est produit. Il est utilisé pour définir le nombre d'agrégats qui sera retenu. Il n'existe pas de règles mathématiques pour la définition de ce nombre. C'est un avis d'expert qui prend en compte la clarté de la séparation entre les groupes et le nombre d'individus présents dans chaque groupe. La démarche peut être itérative, le nombre de groupes étant modifié après l'étape d'interprétation. Dans le cas étudié, six groupes ont été retenus.

Figure 2

Deuxième étape : classification ascendante hiérarchique (CAH)



Dans une troisième étape, les coordonnées moyennes sur chaque axe de l'ACP sont calculées pour chacun des groupes et, à l'aide de la méthode k-means, chaque lot est réaffecté au groupe ayant la moyenne la plus proche. Ceci permet d'optimiser l'homogénéité des groupes. La représentation de ces groupes sur les axes de l'ACP (Figure 3) permet de connaître quels sont les axes qui discriminent les groupes entre eux. Si certains axes s'avèrent non-discriminants une nouvelle itération de l'analyse est réalisée en supprimant les axes non-discriminants de la CAH.

Enfin, dans une quatrième étape les caractéristiques de chaque groupe sont décrites et la signification biologique de ces groupes est recherchée. Dans le cas étudié, la description est visualisée en représentant la distribution des moyennes des titres sérologiques pour chacun des groupes retenus (Figure 4) et par la représentation des titres individuels d'un élevage prototype (Figure **5**). Les élevages prototypes sont, pour chaque groupe, celui qui est le plus proche du barycentre du groupe. Pour ce qui concerne l'interprétation des résultats, une attention particulière est portée au classement des quatre profils réalisés dans des conditions expérimentales. Il est constaté que les deux profils d'animaux non vaccinés sont classés dans le groupe 1 ; on conclut donc que ce groupe comprend les élevages non vaccinés ou très mal vaccinés. Les deux profils de lots vaccinés dans des conditions expérimentales sont dans le groupe 4 ; on conclut donc que ce lot comprend les élevages normalement vaccinés. Le groupe 6 présente la même hiérarchie entre les différentes Elisa que le groupe 4 à un niveau plus élevé ; on pose donc l'hypothèse qu'il s'agit d'élevages très bien protégés ; il est en effet logique que des titres plus élevés que ceux des lots

expérimentaux puissent être obtenus, car seule une primo-vaccination a été réalisée dans les lots expérimentaux, alors que sur le terrain les multipares sont multi-vaccinées si le protocole du fabricant est bien respecté. Les trois autres lots (2, 3 et 5) sont ceux pour lesquels une divergence est observée entre les différentes réponses immunitaires ; ce sont ces lots qui posent problème lors de l'interprétation des profils. Les profils du lot 2 présentent une réponse forte pour deux antigènes (F4ab et F4ac) et faible pour les autres ; on peut remarquer que cela correspond à la discrimination faite par le troisième axe de l'ACP ; après analyse des commémoratifs disponibles et de données bibliographiques, il est posé l'hypothèse qu'il s'agit d'élevages vaccinés avec un autre vaccin, présentant un profil antigénique différent. Enfin, les profils des groupes 3 et 5 présentent des titres supérieurs pour les Elisa F4ac et F5 ; on retrouve là l'information retranscrite dans l'axe 2 de l'ACP ; il n'y a à ce jour pas d'hypothèse pour expliquer ces cas, il pourrait s'agir d'interférences entre les réactions immunitaires post-vaccinales et post-infectieuses ; la conclusion est que l'étude de ces cas doit être approfondie en collectant des informations complémentaires sur le terrain.

Figure 3

Troisième étape : projections des groupes (identifiés par leurs numéros) et des individus

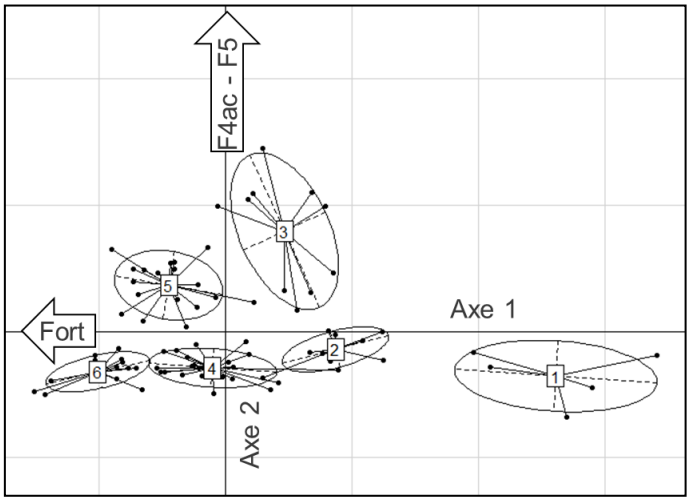


Figure 4

Quatrième étape : description des groupes et interprétation

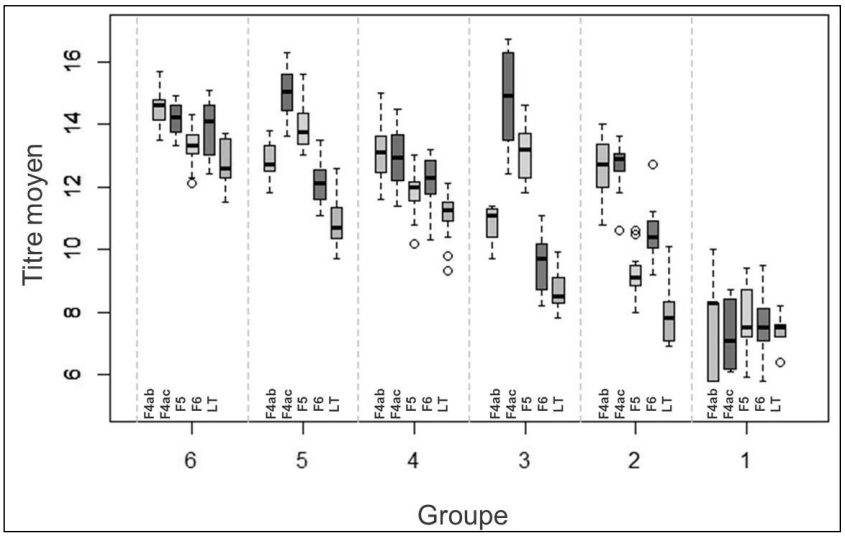
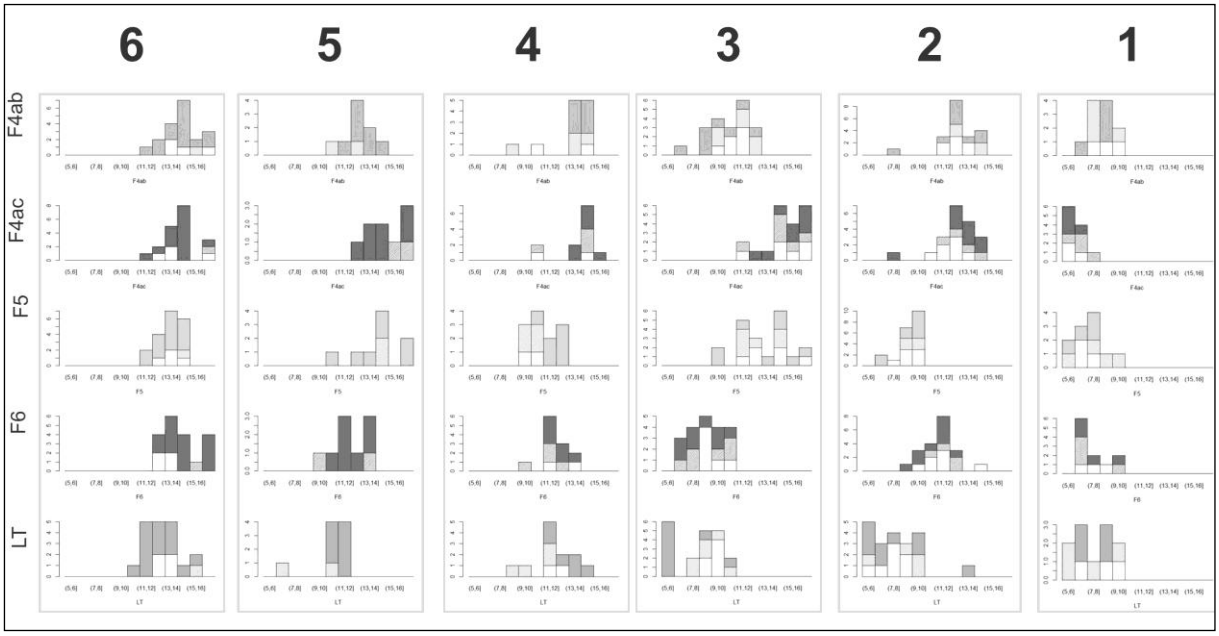


Figure 5

**Représentation des groupes par un élevage prototype :
pour chaque élevage prototype et pour chaque Elisa, la distribution des résultats individuels
(de chaque truie) est donnée**

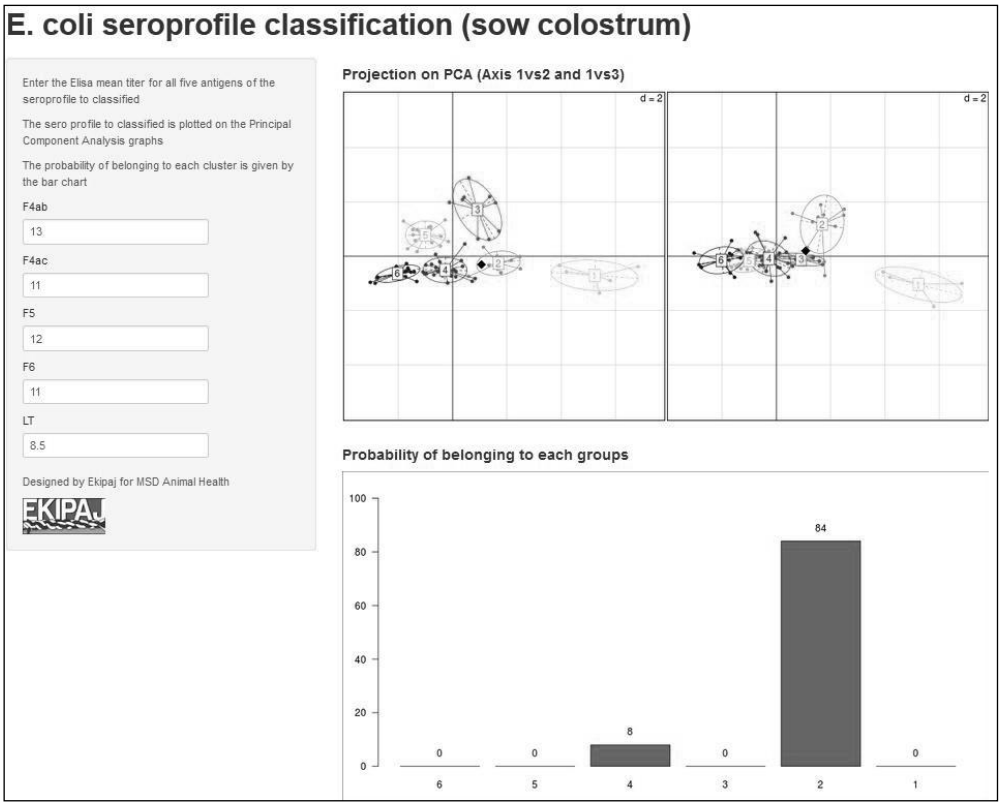


3. UTILISATION DU MODÈLE

La modélisation par typologie est utilisée pour construire un outil d'aide au diagnostic. Il s'agit de déterminer dans quel groupe se situe un élevage nouvellement contrôlé, c'est-à-dire qui n'est pas présent dans la base de données initiale. Le principe est de situer le nouvel élevage dans l'espace décrit pat l'ACP sans relancer la procédure complète. C'est-à-dire que le nouvel élevage ne vient pas modifier les groupes qui ont été constitués précédemment. Ceci est réalisé en calculant la distance entre le nouvel élevage et le barycentre de chacun des groupes puis en calculant une probabilité d'appartenance aux groupes. Cette probabilité est calculée par régression logistique pour chacun des six groupes. Pour cela on réalise un modèle logistique pour chaque groupe, la variable à expliquer étant l'appartenance à ce groupe (0/1) et l'unique variable explicative la distance entre l'individu et le centre du groupe. Dans le cas d'étude, cette procédure a été implémentée en utilisant Shiny pour RStudio [RStudio - Shiny, 2013]. Il s'agit d'une application très simple (un seul écran) accessible par Internet sans installation (Figure 6). Les moyennes des titres sont saisies dans les champs situés dans la partie gauche et les probabilités d'appartenance aux différents groupes sont données dans le graphique en barres. Le nouvel individu est classé dans le groupe pour lequel il a la plus forte probabilité d'appartenance. Si deux groupes ont des probabilités voisines, la classification doit être considérée comme douteuse (il n'a pas encore été défini de seuil précis pour cette notion de douteux). La projection de l'individu, représenté par un carré noir, dans les deux plans de l'ACP (graphiques en haut à droite) peut être utilisée pour affiner l'interprétation.

Figure 6

Copie d'écran de l'application d'aide à la décision



III - DISCUSSION

La typologie est une technique d'analyse de données régulièrement utilisée en épidémiologie humaine et animale. Elle est par exemple de plus en plus utilisée en épidémiologie nutritionnelle car elle permet d'analyser les profils diététiques en prenant en compte les aspects multidimensionnels et multi-colinéaires de ces données. Ceci est un progrès, car les techniques d'analyses plus classiques obligeaient souvent les chercheurs de ce domaine à se limiter à l'exploration des relations entre une maladie chronique et un seul nutriment [Siou *et al.*, 2011]. Dans un domaine voisin, un exemple d'utilisation est une analyse des relations entre asthme et obésité [Sutherland *et al.*, 2012]. En épidémiologie vétérinaire, on trouve des exemples appliqués aux profils d'infection en élevage porcin [Fablet *et al.*, 2012], à la surveillance syndromique à l'abattoir [Dupuy *et al.*, 2013], aux profils de résistance aux antibiotiques [Berge *et al.*, 2003], ou aux profils sérologiques en élevage avicole [Auvigne *et al.*, 2013] . L'analyse typologique peut également être utilisée pour décrire des systèmes d'élevage [Usai *et al.*, 2006].

Cependant, il semble que cette technique soit beaucoup moins utilisée que d'autres méthodes, en particulier les régressions linéaires et régressions logistique. On peut se poser la question des raisons de cette sous-utilisation. Notre hypothèse est que la typologie souffre de présenter des résultats moins tranchés que la régression linéaire, en particulier pour l'épidémiologie analytique. En effet, alors qu'une régression linéaire amènera à écrire que *tel facteur de risque est lié à l'expression de telle maladie à un niveau de risque donné*, quitte à « oublier » les liens entre le facteur de risque retenu et d'autres facteurs de risque, la typologie permettra « seulement » de dire que *l'exposition à tel facteur de risque est plus fréquente chez telle catégorie d'individus, qui présente telle prévalence de maladie*. Cette présentation des résultats est

probablement moins accrocheuse qu'un odds-ratio, mais nous considérons qu'elle a le grand avantage d'inciter l'épidémiologiste à la modestie sur ses conclusions en mettant en avant la multi-colinéarité et en rendant plus difficile un glissement sémantique de la constatation d'une association vers l'inférence causale.

Une autre différence entre typologie et régression logistique est un changement radical quant à l'objet de l'étude. Dans l'exemple que nous avons présenté on passe en effet d'une formulation où une variable est le sujet (l'utilisation de tel vaccin est liée à tel résultat sérologique) à une formulation où l'individu est le sujet (les éleveurs de tel groupe utilisent plus tel vaccin et ont tel résultat sérologique). Cette différence de formulation est très intéressante pour faciliter la communication entre l'épidémiologiste et le vétérinaire clinicien ou entre le vétérinaire clinicien et l'éleveur. En effet, quand un être humain prend des décisions rapides et intuitives (ce qui est le cas pour la majorité de nos décisions), la comparaison entre le cas sur lequel il faut statuer et des prototypes a une grande importance ; « des cas individuels surprenants ont un impact puissant » [Kahneman, 2012]. La typologie présente les résultats sous forme de prototypes (chaque groupe peut être décrit par un prototype). Nous posons donc l'hypothèse qu'elle présente le double avantage d'être adaptée à une intégration dans l'heuristique de jugement du décideur et d'avoir été élaborée à l'issue d'un processus rigoureux de collecte et d'analyse de données.

La rigueur méthodologique de la typologie peut cependant être contestée. Cette méthode peut être considérée comme une méthode de « data-miming » ne nécessitant pas de poser d'hypothèse préalable sur ce qui caractérise la différence entre les groupes. Cette absence d'hypothèse est un élément très intéressant ; cependant, il faut la relativiser. En effet, les résultats obtenus dépendront de toutes les hypothèses, objectifs et contraintes prises en compte lors de la conception du protocole de collecte de données. Dans le cas concret présenté ici, ceci concerne en particulier les modalités d'échantillonnage des élevages et des animaux et le choix des analyses sérologiques. De même, un choix est fait pour déterminer quelles seront les données utilisées pour construire la typologie (ici les résultats des analyses sérologiques) et les données utilisées pour l'interpréter (ici, les commémoratifs sur les schémas vaccinaux). La rigueur méthodologique est donc indispensable, comme pour tout type d'étude.

Enfin, il faut prendre en compte que lors de l'étape de classification (CAH), deux décisions incluant une part importante de subjectivité sont nécessaires : lors du choix du nombre d'axes à inclure, et lors du choix du nombre de groupes à retenir. Des démarches ont été explorées pour limiter cette subjectivité. Il est ainsi proposé un indicateur fondé sur le ratio des variances intra et inter-agrégats pour choisir automatiquement le nombre de groupes [Siou *et al.*, 2011]. Pour notre part, nous préférons assumer cette subjectivité et avoir une démarche itérative. Cette démarche itérative consiste à relancer l'analyse en modifiant les choix de nombre d'axes et d'agrégats, si l'analyse initiale donne des résultats difficilement interprétables.

En conclusion, l'analyse typologique nous apparaît comme une technique adaptée à l'épidémiologie descriptive et analytique, utilisable avec des jeux de données de tout type et particulièrement intéressante pour interagir avec les décideurs et utilisateurs finaux des résultats des études terrain.

BIBLIOGRAPHIE

Auvigne V., Belloc C. - Le clinicien et l'épidémiologiste : deux paradigmes pour un dialogue. *Épidémiol. et santé anim.*, 2012, **61**, 141-147.

Auvigne V., Gibaud S., Léger L., Malher X., Currie R., Riggi A. - A longitudinal study of the incidence of Avian Infectious Bronchitis in France using strain-specific haemagglutination inhibition tests and cluster analysis. *Revue Méd. Vét.*, 2013, (accepté pour publication).

Berge A.C., Atwill E., Sischo W. - Assessing antibiotic resistance in fecal Escherichia coli in young calves using cluster analysis techniques. *Preventive Veterinary Medicine*, 2003, **61** (2), 91-102.

Besson H., Murmans M., Witvliet M. - Compatibility of vaccines against atrophic rhinitis and neonatal E. coli diarrhea. In : Proc. 21th IPVS Congress. Vancouver, Canada, 2010, 88.

Dupuy C., Morignat E., Hendrikx P., Ducrot C., Maugey E., Vinard J., Calavas D., Gay E. - Using bovine meat inspection data for syndromic surveillance: innovative statistical approach for defining syndromes. In : SVEPM meeting proceedings. Madrid, Spain, 2013, 95-104.

Fablet C., Marois-Créhan C., Simon G., Grasland B., Jestin A., Kobisch M., Madec F., Rose N. - Infectious agents associated with respiratory diseases in 125 farrow-to-finish pig herds: a cross-sectional study. *Vet. Microbiol.*, 2012, **157** (1-2), 152-163.

Kahneman D. - Système 1 / Système 2 : Les deux vitesses de la pensée. Flammarion, 2012.

Riising H.-J., Murmans M., Witvliet M. - Protection against neonatal *Escherichia coli* diarrhoea in pigs by vaccination of sows with a new vaccine that contains purified enterotoxic *E. coli* virulence factors F4ac, F4ab, F5 and F6 fimbrial antigens and heat-labile *E. coli* enterotoxin (LT) toxoid. *J. Vet. Med. B Infect. Dis. Vet. Public Health*, 2005, **52** (6), 296-300.

RStudio - Shiny 2013. Available at: http://www.rstudio.com/shiny/ [Accessed February 20, 2013].

Siou G.L., Yasui Y., Csizmadi I., McGregor S.E., Robson P.J. - Exploring Statistical Approaches to Diminish Subjectivity of Cluster Analysis to Derive Dietary Patterns The Tomorrow Project. *Am. J. Epidemiol.*, 2011, **173** (8), 956-967.

Sutherland E.R., Goleva E., King T.S., Lehman E., Stevens A.D., Jackson L.P., Stream A.R., Fahy J.V., Leung D.Y.M. - Cluster analysis of obesity and asthma phenotypes. *PLoS ONE*, 2012, **7** (5), e36631.

Usai M.G., Casu S., Molle G., Decandia M., Ligios S., Carta A. - Using cluster analysis to characterize the goat farming system in Sardinia. *Livestock Science*, 2006, **104** (1-2), 63-76.

Remerciements

Les auteurs remercient MSD Santé Animale, pour l'autorisation d'utilisation des données et Elisabeth Sallé pour l'important travail de terrain réalisé au préalable.