Symbolic Covariance ACP et régression pour variables à valeurs d'intervalles.

Application en épidémiologie vétérinaire.

Stéphanie Bougeard[1], Carole Toque[2,3]

[1] Agence Nationale de Sécurité Sanitaire (Anses), Laboratoire de Ploufragan-Plouzané
[2] Syrokko
[3] Université du Luxembourg

Résumé Cet article positionne l'Analyse de Données Symboliques dans le cadre du traitement statistique des données d'épidémiologie vétérinaire. Une démarche complète d'Analyse de Données Symboliques est illustrée sur un exemple visant à déterminer les facteurs de risque du taux de saisie de poulets de chair à l'abattoir. Les unités statistiques étant les abattoirs dans lesquels plusieurs lots d'animaux sont enquêtés, les variables sont considérées par la suite comme des variables symboliques à valeurs d'intervalles. Deux méthodes sont appliquées : la *Symbolic Covariance PCA* et la *Symbolic Covariance Regression* ; ces méthodes sont basées sur des corrélations et covariances symboliques prenant en compte les deux bornes des intervalles et ayant la propriété d'être décomposables en variations intra- et inter-concepts.
Mots clés : Analyse de données, régression, données symboliques, covariance symbolique, épidémiologie.

1 Introduction

1.1 Contexte

L'épidémiologie vétérinaire consiste en l'étude des maladies dans une population animale. La principale étape, *i.e.*, l'épidémiologie analytique, vise à déterminer les facteurs de risque liés à l'apparition et au développement d'une maladie. Selon la connaissance de celle-ci par l'épidémiologiste, l'unité statistique peut être l'élevage ou l'animal. Le protocole, les traitements statistiques ainsi que les conclusions, sont associés à cette unité. La majorité des variables est recueillie au niveau des animaux si l'unité est l'animal, ou des élevages si l'unité est l'élevage. Si l'unité est l'animal, l'enquête est basée sur un nombre limité d'élevages dans lesquels un nombre représentatif d'animaux est tiré au sort. Si l'unité est l'élevage, l'enquête est basée sur un nombre représentatif d'élevages sélectionnés par tirage au sort. Si l'unité statistique est l'élevage, la majorité des variables est récoltée sur ceux-ci (*e.g.*, durée du vide sanitaire, taux d'ammoniac), mais quelques mesures peuvent être réalisées sur les animaux (*e.g.*, poids, portage de virus). Les études d'épidémiologie sont basées sur des questionnaires objectivant la structure et l'environnement des élevages et sur des mesures sur les animaux. Il s'ensuit que la base de données de l'enquête est structurée en nombreuses thématiques, comme les caractéristiques de l'élevage (nombre d'animaux, performances zootechniques, ...) ou l'état sanitaire du troupeau (dosages sérologiques, pesées, traitements antibiotiques, ...). Les variables sont mesurées soit une seule fois (*e.g.*, taille de l'élevage) soit plusieurs fois au cours du temps (*e.g.*, poids des