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

animaux). Du fait de la structure complexe des observations, la base de données associée à une enquête d'épidémiologie est constituée de différents sous-tableaux mesurés sur des unités différentes.

1.2 Problématique et analyses statistiques standards

La principale question à laquelle l'épidémiologiste doit répondre est la détermination de facteurs de risque de la maladie étudiée. La problématique statistique globale est donc le plus souvent la régression. A celle-ci, s'ajoutent quatre principales contraintes qui posent une problématique de régression complexe. [C1] La première contrainte est relative à la structure des observations qui diffère selon les variables, *e.g.*, mesures au niveau de l'animal, de l'élevage, parfois répétées au cours du temps. Lorsque l'unité statistique est l'animal, la régression cherche à s'affranchir des différences entre élevages. Lorsque l'unité est l'élevage, la régression considère les animaux comme des répétitions de l'unité étudiée. [C2] La seconde contrainte est relative aux nombreuses variables explicatives à inclure dans la régression, celles-ci présentant de plus des quasi-colinéarités marquées. Par ailleurs, ces variables sont aussi organisées en blocs ayant un sens biologique, *e.g.*, variables relatives à l'hygiène ou aux pratiques zootechniques, dont il est intéressant de tenir compte. [C3] La troisième contrainte tient au fait que la maladie étudiée est caractérisée par plusieurs mesures, *e.g.*, signes cliniques, résultats relatifs aux portages de virus. La prise en compte de ces différentes composantes du diagnostic enrichit l'interprétation, en comparaison aux qualifications usuelles (malade *vs* non malade). [C4] La dernière contrainte est relative à la nature variée des données : qualitatives (booléenne, ordinale, nominale) et quantitatives associées à différentes distributions (uniforme, normale, Poisson, ...).

Selon que l'unité statistique est l'animal ou l'élevage, la façon de conduire le traitement statistique diffère. (i) Lorsque l'unité est l'animal, la majorité des variables est mesurée à ce niveau, les variables mesurées au niveau de l'élevage étant dupliquées pour les animaux lui appartenant. Les facteurs de risque sont recherchés au niveau de l'animal en cherchant à s'affranchir des différences entre élevages. Afin de pallier les contraintes C2 et C3, la régression est précédée d'analyses factorielles supervisées. L'effet structurant de l'élevage devant être écarté (C1), des analyses intra-groupes (ou multigroupes), le groupe étant l'élevage, sont appliquées. Une fois les variables sélectionnées et/ou synthétisées, un modèle de régression prenant en compte la nature variée des données (C4) est établi par modèle linéaire généralisé marginal ou équation d'estimation généralisée [22], ceux-ci corrigeant les estimateurs pour intégrer la non-indépendance des observations. (ii) Lorsque l'unité est l'élevage, la majorité des variables est mesurée à ce niveau, les variables mesurées au niveau de l'animal étant synthétisées, *e.g.*, médiane pour les variables quantitatives, fréquence de positifs pour les variables booléennes. Les facteurs de risque sont recherchés au niveau de l'élevage. Comme précédemment du fait des contraintes C2 et C3, des analyses factorielles (multiblocs) supervisées sont utilisées pour sélectionner et/ou créer des variables de synthèse. Par la suite, un modèle linéaire généralisé classique [27], adapté à la nature des variables (C4), est établi.

1.3 Apports de l'analyse de données symboliques

Lorsque l'unité statistique est l'élevage, la majorité des mesures est réalisée dans cette unité (*e.g.*, taille des bâtiments), mais d'autres en sont des répétitions temporelles (*e.g.*,

taux d'ammoniac au cours d'une journée) ou des mesures répétées sur plusieurs animaux (*e.g.*, portage du virus grippal). Les élevages peuvent être considérés comme des concepts, *i.e.*, observations de second ordre sur lesquels l'analyse est centrée, et les animaux ou rép étitions temporelles comme des instances de concepts, *i.e.*, observations de premier ordre d écrivant les variations internes de ces concepts. Les élevages sont décrits par des variables de différents types : valeurs uniques qualitatives ou quantitatives (mesures au niveau de l'élevage) ou multiples pondérées de type diagramme (mesures qualitatives tem-porelles ou sur les animaux) ou histogramme (mesures quantitatives temporelles ou sur les animaux, éventuellement recodées sous forme d'intervalles). Ces variables de différents types sont regroupées sous l'appellation de variables symboliques. La description des indi-vidus de second ordre par ces variables réduit les pertes d'information due à l'utilisation de fr équences ou médianes. L'analyse de données symboliques étend les méthodes standards aux variables symboliques. La régression multiple symbolique, par exemple, généralise la régression aux concepts présentant à la fois des valeurs uniques, multiples (intervalles) et multiples pondérées (histogrammes) [3, 4, 1] ; comme cela va être montré par la suite sur nos données de type intervalle, elle solutionne en partie la contrainte C1 relative aux différentes unités statistiques. Cependant, cette méthode reste sensible aux mêmes limites que la régression standard, *i.e.*, non-stabilité des coefficients de régression lors de l'introduction de nombreuses variables explicatives quasi-colinéaires (C2). Il convient donc de la précéder d'une analyse factorielle symbolique [10, 4, 20, 26] pour sélectionner les variables et/ou les synthétiser (C2 et C3). Cette approche est détaillée (Section 2) et illustrée (Section 3) sur des données d'épidémiologie à valeurs d'intervalles.

2 Méthodes

2.1 Données et objectifs

Les données proviennent d'une enquête d'épidémiologie visant à déterminer les facteurs de risque de la mortalité au cours de l'élevage ainsi que des réformes à l'abattoir pour la filière poulet de chair [25]. Elle a été menée en 2005 par l'Agence Nationale de Sécurité Sanitaire dans quinze abattoirs nationaux. Une cohorte de 324 lots de poulets est échantillonnée selon le volume et le jour d'abattage (entre 10 et 35 lots par abattoir). Pour chacun de ces lots, une enquête prospective est menée à l'abattoir consistant à enregistrer les conditions de transport et d'abattage ainsi que le taux de saisie par les Services Vétérinaires. Lors d'une enquête rétrospective, les conditions d'élevage, l'historique de santé, la mortalité journalière ainsi que les conditions de ramassage de ces mêmes lots ont été collectés. L'objectif est d'expliquer le taux de saisie à l'abattoir (Condemn) par une sélection de neuf variables : le taux de mortalité la première semaine (Mort7), durant la période d'élevage (Mort) et durant le transport (Doa), la surface d'élevage des animaux (Area), le nombre de poussins à la mise en place (NbChick), la fréquence de visite de l'éleveur durant l'élevage (Freqchicken), la densité des animaux dans les caisses de transport (StockingD), la distance entre l'élevage et l'abattoir (DistSlaught) et le temps d'attente à l'abattoir (Tlairage). Les unités statistiques considérées sont les abattoirs ; en effet, il est plus intéressant de considérer ceux-ci plutôt que les lots, car les abattoirs ont des pratiques différentes, notamment en termes de saisie, et sont associés à des mêmes groupements de producteurs ayant pratiques d'élevage similaires. En considérant

les abattoirs comme des concepts et les lots comme des instances de concepts mesurant les variations de ceux-ci, les variables symboliques à valeurs d'intervalles sont créées (Table 1). Par la suite, les variables standards sont notées en minuscule et les variables symboliques correspondantes en majuscule.

TABLE 1 – Données symboliques relatives à l'étude des facteurs de risque du taux de saisie des poulets de chair à l'abattoir. Les abattoirs sont décrits par des variables symboliques à valeurs d'intervalles (*e.g.*, valeurs minimales et maximales de l'intervalle).

Concepts	Instances	CONDEMN	MORT7	MORT	DOA	AREA	...
Abattoir 1	N=16	[0.6 ;2.1]	[0.2 ;2.2]	[0.6 ;2.7]	[0 ;0.4]	[1000 ;5200]	...
Abattoir 2	N=15	[0.2 ;0.9]	[0.4 ;2.2]	[0.8 ;4.1]	[0 ;0.9]	[1000 ;4000]	...
Abattoir 3	N=18	[0.2 ;2.1]	[0.4 ;2.0]	[0.2 ;2.9]	[0 ;0.9]	[1000 ;5000]	...
...	...	...	...	...	...	...	...
Abattoir 15	N=10	[0.1 ;0.9]	[1.0 ;2.8]	[0.5 ;2.3]	[0.1 ;0.6]	[969 ;2600]	...

Concepts	...	NBCHICK	FREQCHICKEN	TLAIRAGE	STOCKINGD	DISTSLAUGHT
Abattoir 1	...	[17340 ;32946]	[1 ;6]	[105 ;240]	[49 ;71]	[18 ;129]
Abattoir 2	...	[10710 ;34990]	[1 ;5]	[45 ;195]	[41 ;71]	[3 ;102]
Abattoir 3	...	[10506 ;39780]	[1 ;6]	[35 ;250]	[42 ;58]	[8 ;133]
...	...	...	...	...	...	...
Abattoir 15	...	[10400 ;37000]	[2 ;4]	[105 ;300]	[50 ;68]	[4 ;215]

2.2 Statistiques pour variables à valeurs d'intervalles

Le traitement statistique de variables à valeurs d'intervalles est classiquement basé sur la transformation de ces intervalles en leurs centres ou en leurs étendues (ou rangs). Les statistiques classiques, *e.g.*, analyse en composantes principales ou régression, peuvent ensuite être appliquées directement sur ces variables simplifiées. Cependant, l'information contenue dans ces nouvelles variables est réduite et leur variation totale est mal prise en compte.

Afin de tenir compte au mieux du format des variables à valeurs d'intervalles, de nouvelles statistiques sont définies à partir des fonctions de densité de variables aléatoires, sous l'hypothèse que celles-ci suivent une loi uniforme sur ces intervalles. De façon plus formelle, soit deux variables quantitatives X_i et X_j basées sur N concepts à valeurs d'intervalles, définies par $X_{ni} = [a_{ni}, b_{ni}]$ et $X_{nj} = [a_{nj}, b_{nj}]$ pour $n = [1, \ldots, N]$. En supposant que la distribution de X_{ni}, resp. X_{nj}, suit une loi uniforme dans l'intervalle $[a_{ni}, b_{ni}]$, resp. $[a_{nj}, b_{nj}]$, pour chacun des $n = (1, \ldots, N)$ concepts considérés, sont définies leur moyenne par l'Eq. (1), leur variance par l'Eq. (2), leur covariance par l'Eq. (3) et leur corrélation par l'Eq. (4). La justification ainsi que les détails de ces calculs sont donnés par Bertrand [2] pour le cas univarié et par Billard [6, 7] pour le cas bivarié.

$$\bar{X}_i = \frac{1}{2N}\sum_n (a_{ni} + b_{ni}) \tag{1}$$

$$var(X_i) = \frac{1}{3N}\sum_n (a_{ni}^2 + a_{ni}b_{ni} + b_{ni}^2) - \frac{1}{4N^2}\left[\sum_n (a_{ni} + b_{ni})\right]^2 \tag{2}$$

$$cov(X_i, X_j) = \frac{1}{3N}\sum_n G_{ni}G_{nj}\sqrt{Q_{ni}Q_{nj}} \tag{3}$$

$$\text{avec} \quad Q_{ni} = (a_{ni} - \bar{X}_i)^2 + (a_{ni} - \bar{X}_i)(b_{ni} - \bar{X}_i) + (b_{ni} - \bar{X}_i)^2$$

$$\text{et} \quad G_{ni} = \begin{cases} -1 & \text{si} \quad (a_{ni} + b_{ni})/2 \leq \bar{X}_i \\ 1 & \text{si} \quad (a_{ni} + b_{ni})/2 > \bar{X}_i \end{cases}$$

$$cor(X_i, X_j) = \frac{cov(X_i, X_j)}{\sqrt{var(X_i).var(X_j)}} \tag{4}$$

L'intérêt de ces formulations est que la variation totale est décomposable en variation inter et intra-concept. Pour le cas univarié, Billard [6, 7] démontre que la variation totale d'une variable à valeurs d'intervalles, *i.e.*, $var(X_i)$, se décompose en variation inter-concept et intra-concept, ces variations étant calculées à partir du centre de chacun des N intervalles $(a_{ni}+b_{ni})/2$ pour $n = (1, \ldots, N)$. Pour le cas bivarié, Billard [7] démontre que la covariance entre deux variables à valeurs d'intervalles, *i.e.*, $cov(X_i, X_j)$, se décompose en co-variation inter et intra-concept. Les méthodes utilisées par la suite, *i.e.*, ACP et régression, sont respectivement basés sur les matrices de corrélation et de variance-covariance définies à partir de ces équations.

2.3 *Symbolic Covariance PCA*

Plusieurs méthodes d'ACP pour variables à valeurs d'intervalles sont proposées dans la littérature : la méthode des centres [10, 11, 6], des sommets [10, 11, 6], les méthodes Symbolic-Object PCA et Range-Transformation PCA pouvant être couplées [17], la méthode des centres et des rayons proposée par [28] ou encore la méthode interval-PCA [14]. Toutes ces méthodes sont détaillées dans [20, 21, 18] et comparées dans [18]. Elles présentent des limites soit en termes de structure de la covariance, soit en termes de représentation. De façon plus détaillée, la méthode des centres ne prend en compte que la variation inter-concept. La méthode des sommets où, pour chaque concept, toutes les combinaisons des bornes des intervalles sont données, considère improprement ceux-ci comme des informations indépendantes. De plus, ces méthodes proposent des représentations graphiques des concepts à valeurs d'intervalles par des rectangles MCAR (=Maximum Covering Area Rectangles, rectangle minimal de l'enveloppe convexe issue de la projection du polytope) ou des rectangles tronqués PECS (=Parallel Edge Connected Shapes, polygone (*i.e.*, rec-tangle MCAR tronqué) qui couvre au plus près les points de la projection du polytope, proposé par [15]). Ces deux représentations définissent des zones trop larges où les données peuvent ne pas être présentes.

Soit P variables $\mathbf{X} = (X_1, \ldots, X_P)$ formées par N concepts et définies par des intervalles. Nous proposons d'appliquer une ACP basée sur la décomposition en valeurs spectrales de la matrice de corrélation symbolique $[cor(X_i, X_j)]_{P\times P}$ définie à partir de l'Eq. (4).

Cette méthode, appelée *Symbolic Covariance PCA*, est proposée par Le-Rademacher [18]. La décomposition en valeurs spectrales de la matrice de corrélation symbolique donne le vecteur des valeurs propres $\Lambda = (\lambda^{(1)}, \ldots, \lambda^{(H)})$ et la matrice des vecteurs propres $\mathbf{W} = [w^{(1)}, \ldots, w^{(H)}]$ selon l'Eq. (5), avec $H = rang([cor(X_i, X_j)]_{P \times P})$.

$$[cor(X_i, X_j)]_{P \times P}.w^{(h)} = \lambda^{(h)} w^{(h)} \quad \text{pour} \quad h = (1, \ldots, H) \tag{5}$$

Le vecteur des valeurs propres Λ fournit les pourcentages d'inertie de chaque dimension ; la matrice des vecteurs propres $\mathbf{W}$ permet la projection des variables dans l'espace des concepts. Les N concepts étant définis par des intervalles, la représentation de ceux-ci dans l'espace des variables est définie par N hyper-rectangles reflétant leur structure réelle. Les frontières de chaque hyper-rectangle dans l'espace des P variables sont construites à partir de leur matrice des sommets [18] ; cette matrice $\mathbf{X}_n^{\nu}$ de dimension $(2^P \times P)$ est donnée par l'Eq. (6) pour P $(p = 1, \ldots, P)$ variables à valeurs d'intervalles $X_{np} = [a_{np}, b_{np}]$ et un concept n donné.

$$\mathbf{X}_n^{\nu} = \begin{pmatrix} a_{n1} & a_{n2} & \ldots & a_{nP} \\ a_{n1} & a_{n2} & \ldots & b_{nP} \\ \vdots & \vdots & \vdots & \vdots \\ b_{n1} & b_{n2} & \ldots & a_{nP} \\ b_{n1} & b_{n2} & \ldots & b_{nP} \end{pmatrix} \quad \text{pour chaque } n = (1, \ldots, N) \tag{6}$$

La projection des N hyper-rectangles dans l'espace des vecteurs propres $\mathbf{W} = [w^{(1)}, \ldots, w^{(H)}]$ de la *Symbolic Covariance PCA* donne N polytopes, issus de la transformation linéaire des matrices de sommets selon l'Eq. (7). Chacune de ces N matrices $(\mathbf{T}_1, \ldots, \mathbf{T}_N)$ est de dimension $(2^P \times H)$.

$$\mathbf{T}_n = \mathbf{X}_n^{\nu} \mathbf{W} \quad \text{pour chaque } n = (1, \ldots, N) \tag{7}$$

Conformément aux objectifs donnés paragraphe 2.1, nous souhaitons une représentation des variables actives $\mathbf{X} = (X_1, \ldots, X_P)$ associée à la projection d'une variable supplémentaire (à expliquer) Y dans le cadre de la *Symbolic Covariance PCA*. Cette problématique n'a pas encore été traitée en Analyse de Données Symboliques. Pour cela, nous proposons une représentation des variables à partir des corrélations entre les variables (actives et supplémentaire) et les composantes symboliques, en remplacement à la représentation associée aux vecteurs propres $\mathbf{W}$. Afin de calculer ces corrélations selon l'Eq. (4), les N polytopes sont remplacés par N intervalles ; géométriquement, ceci revient à simplifier les polytopes par des rectangles MCAR. Il s'ensuit une matrice de composantes à valeurs d'intervalles de dimension $(N \times H)$ dont les bornes sont données par l'Eq. (8) pour chaque concept n et chaque dimension h.

$$\begin{aligned} CP.inf_n^{(h)} &= min(\mathbf{T}_n^{(h)}) \quad \text{pour } n = (1, \ldots, N),\ h = (1, \ldots, H) \\ CP.sup_n^{(h)} &= max(\mathbf{T}_n^{(h)}) \quad \text{pour } n = (1, \ldots, N),\ h = (1, \ldots, H) \end{aligned} \tag{8}$$

Pour une dimension h donnée, les bornes des composantes symboliques à valeurs d'intervalles $CP^{(h)}$ sont fournies par les vecteurs $CP.inf^{(h)} = [CP.inf_1^{(h)}, \ldots, CP.inf_N^{(h)}]$ et $CP.sup^{(h)} = [CP.sup_1^{(h)}, \ldots, CP.sup_N^{(h)}]$. La représentation du cercle des corrélations proposée par la suite est issue des corrélations symboliques entre les matrices $(X_1, \ldots, X_P, Y)$ et $(CP^{(1)}, \ldots, CP^{(H)})$ calculées à partir de l'Eq. (4).

2.4 *Symbolic Covariance Regression*

Plusieurs méthodes de régression pour variables à valeurs d'intervalles sont proposées dans la littérature : la régression sur les centres des intervalles [3], la régression sur les centres et rangs, où ceux-ci sont pris en compte par deux modèles indépendants [9], et la régression bivariée où centres et rangs sont pris en compte conjointement dans un même modèle [6, 23]. Ces méthodes présentent certaines limites ; la méthode des centres ne prend en compte que les variations inter-concepts des données et la méthode des rangs est basée sur une simplification de la variation intra-concept réelle. De ce fait, les méthodes des centres et des rangs ainsi que bivariée ne prennent pas correctement en compte la variation totale des données. Par ailleurs, la méthode des centres et des rangs est basée sur deux régressions indépendantes, ce qui complexifie l'interprétation de la significativité des variables ; la méthode bivariée double le nombre de variables, en supposant de plus leur indépendance, ce qui pose des problèmes de puissance statistique.

Nous proposons d'appliquer une régression dont l'estimateur des moindres carrés est basé sur des matrices de variance-covariance symboliques définies à partir de l'Eq. (3). Cette méthode, proposée par Xu [31], est appelée par la suite *Symbolic Covariance Regression* (=SCR). De façon plus formelle, soit P variables explicatives $\mathbf{X} = (X_1, \ldots, X_P)$ et une variable à expliquer Y, ces variables sont définies par les intervalles $X_{np} = [a_{np}, b_{np}]$ pour $p = (1, \ldots, P)$ et $Y_n = [c_n, d_n]$ pour chacun des $n = (1, \ldots, N)$ concepts. En utilisant les matrices d'intervalles centrées, Xu [31] démontre que l'estimateur des moindres carrés du vecteur des coefficients de régression $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_P)$ est donné par l'Eq. (9), en utilisant l'Eq. (3) pour le calcul des covariances entre ces variables à valeurs d'intervalles. La constante du modèle est donnée par l'Eq. (10) en utilisant l'Eq. (2) pour la valeur moyenne de la variable à expliquer.

$$\hat{\beta} = [Ncov(X_i, X_j)]^{-1}_{P\times P}[Ncov(X_i, Y)]_{P\times 1} \quad \text{avec} \quad i = j = (1, \ldots, P) \tag{9}$$

$$\hat{\beta}_0 = \bar{Y} - \bar{X}\hat{\beta} \tag{10}$$

Bien que l'on soit en présence de variables $\mathbf{Y}$ et $\mathbf{X}$ à valeurs d'intervalles, il est intéressant de noter que, comme la constante $\hat{\beta}_0$ et le vecteur des coefficients de régression $\hat{\beta}$ tiennent à la fois compte des bornes inférieures et supérieures des intervalles dans leurs calculs, ils sont donnés sous forme de réels et non d'intervalles pour chaque association entre variables explicatives et à expliquer ; en comparaison aux méthodes pré-citées pour lesquelles les coefficients de régression sont multiples pour un modèle donné, la méthode SCR propose donc une interprétation plus aisée des liens entre variables.

Dans l'objectif d'évaluer la qualité du modèle de régression, la prédiction de la variable à expliquer ainsi que les résidus et erreurs de prédiction, sont définis. Il convient de noter que ce modèle de régression fournit des estimations du vecteur des bornes inférieures $[\hat{c}_1, \ldots, \hat{c}_N]$, resp. supérieures $[\hat{d}_1, \ldots, \hat{d}_N]$, de la variable à expliquer à valeurs d'intervalles, à partir de la matrice des bornes inférieures $\mathbf{A} = [a_{np}]_{N\times P}$, resp. supérieures $\mathbf{B} = [b_{np}]_{N\times P}$, des P variables explicatives. Ces estimations des deux bornes sont calculées par la même constante $\hat{\beta}_0$ et le même vecteur de coefficients $\hat{\beta}$ définis Eqs. (9) et (10). Afin d'éviter, lors des prédictions, que les bornes inférieures ou supérieures des intervalles ne se trouvent inversées, ce sont en fait les valeurs minima, resp. maxima, de prédiction qui sont affectées au vecteur $[\hat{c}_1, \ldots, \hat{c}_N]$, resp. $[\hat{d}_1, \ldots, \hat{d}_N]$, pour chaque concept $n = (1, \ldots, N)$. De façon

plus formelle, soit les deux vecteurs $Y_{inf} = [c_1, \ldots, c_N]$ et $Y_{sup} = [d_1, \ldots, d_N]$. La prédiction de ces deux vecteurs à partir du modèle précédent $(\hat{\beta}_0, \hat{\beta})$ est fournie par l'Eq. (11).

$$\hat{Y}_{inf} = min(\hat{\beta}_0 + \mathbf{X}\hat{\beta}) \quad \text{et} \quad \hat{Y}_{sup} = max(\hat{\beta}_0 + \mathbf{X}\hat{\beta}) \tag{11}$$

Il s'ensuit le calcul de la matrice des résidus du modèle $\mathbf{RES} = (RES_{inf}, RES_{sup})$, donnés par l'Eq. (12).

$$RES_{inf} = \hat{Y}_{inf} - Y_{inf} \quad \text{et} \quad RES_{sup} = \hat{Y}_{sup} - Y_{sup} \tag{12}$$

Il est ainsi possible de calculer les erreurs de prédiction $\mathbf{RMSE} = (RMSE_{inf}, RMSE_{sup})$ par l'Eq. (13).

$$RMSE_{inf} = \sqrt{\frac{\sum_n RES_{inf}^2}{N}} \quad \text{et} \quad RMSE_{sup} = \sqrt{\frac{\sum_n RES_{sup}^2}{N}} \tag{13}$$

3 Application aux données d'épidémiologie vétérinaire

3.1 Statistiques pour variables à valeurs d'intervalles

Les données initiales (N_i=324 lots répartis dans N_c=15 abattoirs) sont transformées en données symboliques à valeurs d'intervalles, les concepts étant les abattoirs, par la fonction `classic.to.sym` du package `R RSDA` [29]. Afin d'appliquer les calculs et méthodes présentées précédemment, il convient de vérifier que les données de chaque intervalle, associé à chaque variable et chaque concept, suivent une loi de distribution uniforme. Cette hypothèse est vérifiée pour 56% des intervalles (test de Kolmogorov-Smirnov appliqué grâce à la fonction `ks.test` du logiciel `R` au seuil de 1%).

En préalable à l'ACP et à la régression linéaire, les corrélations symboliques pour variables à valeurs d'intervalles peuvent être calculées par la fonction `sym.cor` du même package, qui propose plusieurs options : `centers`, `interval`, `billard`, `histogram`. L'option `billard` utilisée par la suite fournit les covariances et corrélations symboliques données par les formules (3) et (4). Les résultats sont fournis dans la Table 2.

TABLE 2 – Matrice de corrélations symboliques pour variables à valeurs d'intervalles. Données relatives aux saisies de poulets de chair à l'abattoir (N_c=15 concepts).

	MORT7	MORT	DOA	CONDEMN	AREA	NBCHICK	FREQCHICKEN	TLAIRAGE	STOCKINGD
MORT	**-0.38**								
DOA	0.13	-0.02							
CONDEMN	-0.02	-0.12	-0.10						
AREA	**0.38**	-0.26	0.19	**0.47**					
NBCHICK	**0.48**	-0.22	0.02	**0.32**	**0.63**				
FREQCHICKEN	-0.04	-0.20	0.24	0.22	0.05	0.26			
TLAIRAGE	**0.33**	0.07	-0.12	-0.24	0.17	0.15	**-0.40**		
STOCKINGD	-0.04	-0.26	-0.04	-0.16	0.26	0.08	-0.04	0.03	
DISTSLAUGHT	0.11	**0.31**	-0.04	-0.16	0.23	**0.41**	**-0.36**	**0.71**	0.29

Il s'ensuit que certaines corrélations marquées apparaissent, comme entre la surface d'élevage des animaux (AREA) et le nombre de poussins à la mise en place (NBCHICK), ou entre la distance entre l'élevage et l'abattoir (DISTSLAUGHT) et le temps d'attente

sur le quai de l'abattoir (TLAIRAGE). Les corrélations les plus intéressantes concernent la variable à expliquer (CONDEMN) et les deux variables (corrélées) de surface d'élevage (AREA) et nombre de poussins (NBCHICK). Il est possible de les illustrer grâce à la fonction `sym.scatterplot` du package `RSDA`. Les graphes de corrélations entre variables à valeurs d'intervalles sont donnés par la Figure 1.

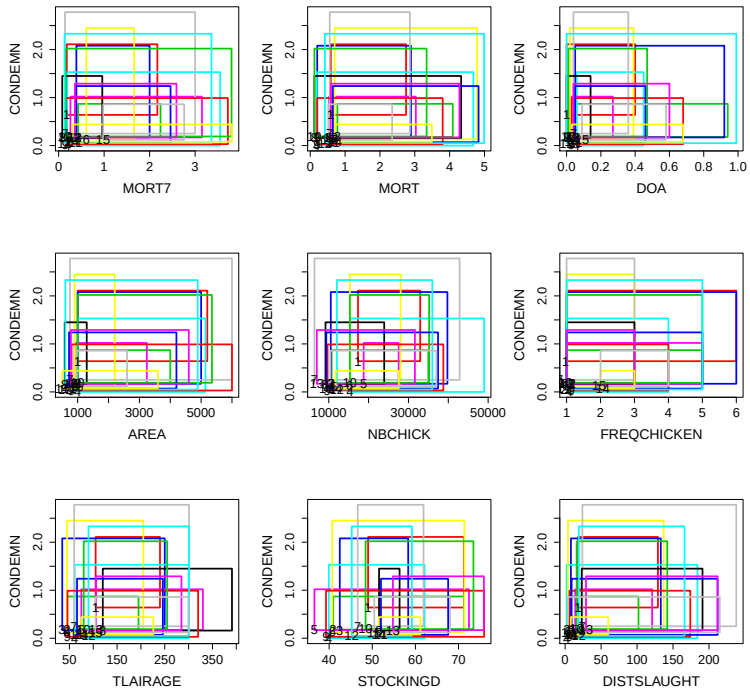


FIGURE 1 – Représentation des corrélations entre variables à valeurs d'intervalles. Données relatives aux saisies de poulets de chair à l'abattoir (N_c=15 concepts).

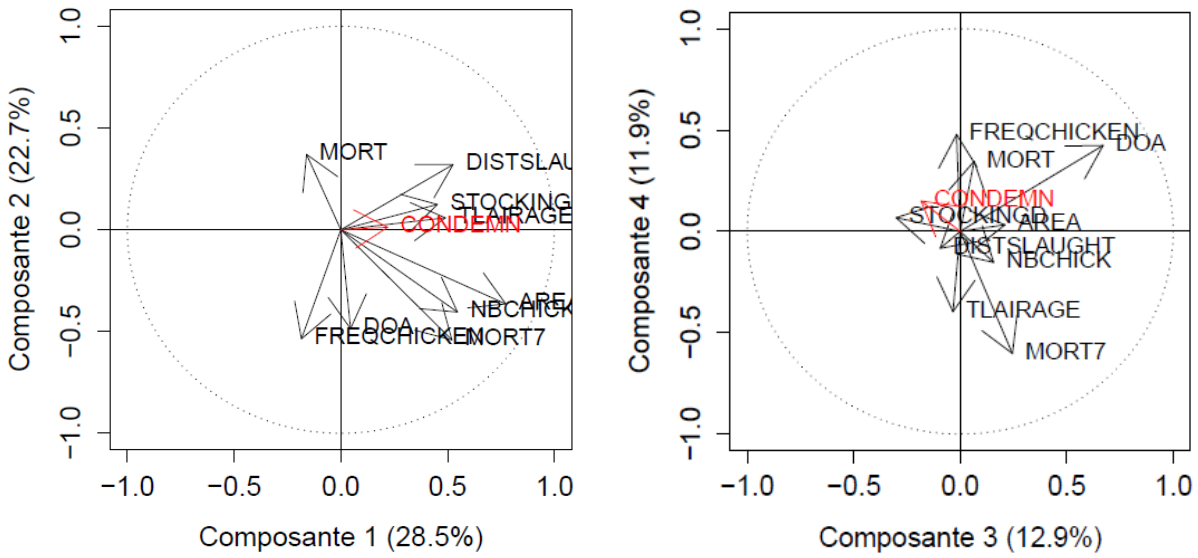
3.2 Résultats de la *Symbolic Covariance PCA*

La méthode *Symbolic Covariance PCA* développée pour variables à valeurs d'intervalles, est conduite au préalable à la régression pour comprendre les liens entre variables et écarter, dans un souci de parcimonie et d'auto-corrélation, les variables ou groupes de variables explicatives les plus corrélées de celle-ci. Comme expliqué au paragraphe 2.3, la *Symbolic Covariance PCA* est basée sur la matrice des corrélations symboliques entre les variables actives. Il s'ensuit que la représentation des variables symboliques est affranchie de leurs échelles de notation, différentes dans cet exemple. La variable à expliquer CONDEMN, qui a statut particulier par rapport aux neuf variables explicatives, est considérée en tant que variable supplémentaire. Les pourcentages d'inertie, qui permettent de sélectionner le nombre de dimensions à interpréter, sont donnés dans la Table 3.

L'interprétation de quatre dimensions de l'espace couvre 76.0% de l'inertie. Les cercles de corrélations des dimensions 1-2 et 3-4 sont donnés en Figure 2. Ces graphes sont associés à des représentations factorielles des concepts par des polytopes (non données ici).

TABLE 3 – Pourcentage d'inertie de chaque dimension de la *Symbolic Covariance PCA*. Données relatives aux saisies de poulets de chair à l'abattoir (N_c=15 concepts).

Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8	Dim.9	Total
28.5%	22.7%	12.9%	11.9%	9.8%	6.4%	4.3%	3.1%	0.45%	100%



(a) Cercle des corrélations (dim. 1-2)

(b) Cercle des corrélations (dim. 3-4)

FIGURE 2 – Cercles des corrélations de la *Symbolic Covariance PCA*. Données relatives aux saisies de poulets de chair à l'abattoir (N_c=15 concepts).

L'interprétation des résultats de la *Symbolic Covariance PCA* poursuit deux objectifs : interpréter les liens entre variables et sélectionner les variables les plus intéressantes pour l'étape suivante de régression. Concernant ce dernier objectif, les variables sont sélectionnées selon deux critères : une corrélation réduite avec les autres variables explicatives, ainsi qu'une corrélation optimale avec la variable à expliquer. Il convient de noter que la variable à expliquer COMDEMN est assez mal projetée sur le plan des variables actives. Les variables AREA, NBCHICK et MORT7 sont liées entre elles sur plan 1-2. Il convient donc de ne sélectionner que l'une d'entre elles dans la régression symbolique visant à expliquer la variable CONDEMN. Parmi ces trois variables, la variable AREA est sélectionnée car elle est la plus liée à la variable à expliquer. La variable DOA est liée à la variable FREQCHICKEN sur le plan 1-2 et est de plus peu liée à la variable à expliquer COMDEMN ; elle n'est pas non plus sélectionnée pour la régression symbolique. La variable DISTSLAUGHT présente une corrélation marquée avec la variable TLAIRAGE sur le plan 1-2 notamment, et est moyennement liée à la variable à expliquer COMDEMN ; elle n'est pas non plus sélectionnée.

3.3 Résultats de la *Symbolic Covariance Regression*

La *Symbolic Covariance Regression* (SCR) pour variables à valeurs d'intervalles, proposée au paragraphe 2.4, est appliquée pour expliquer la variable CONDEMN par les cinq variables explicatives sélectionnées par la *Symbolic Covariance PCA* : MORT, AREA, FRE-

QCHICKEN, TLAIRAGE et STOCKINGD. Cette méthode est programmée sur le logiciel `R`. Il est intéressant de comparer cette méthode à des méthodes plus classiques de régression pour variables à valeurs d'intervalles référencées dans le paragraphe 2.4 : la méthode des centres (fonction `sym.lm` avec l'option "`cm`" du package `RSDA`) et la méthode des centres et des rangs (fonction `sym.lm` avec l'option "`crm`" du package `RSDA`). Il serait intéressant d'inclure dans cette comparaison la méthode bivariée (fonction `bivar` du package `iRegression`) ; cependant, celle-ci a pour variables explicatives le tableau concaténé des centres et des étendues associés à ces variables. Le doublement du nombre de variables explicatives n'est ici pas compatible avec le nombre de concepts du jeu de données (*i.e.*, N_c=15 abattoirs), notamment lors de la validation croisée détaillée ci-après. En effet, la performance des régressions symboliques est évaluée par validation croisée où les (N_c=15) concepts sont divisés 500 fois en données de calibration (2/3) et de validation (1/3). Les régressions symboliques sont ensuite comparées sur la base de deux indices : la corrélation entre les valeurs de la variable à expliquer observée et prédite en utilisant la fonction `sym.cor` du package `RSDA`, ainsi que l'erreur de prédiction (RMSE) dont la formule est donnée par l'Eq. (13), ces deux indices étant calculés pour chacune des 500 données de calibration et de validation. Ces deux résultats sont donnés par la Figure 3.

Les résultats des trois régressions symboliques sont comparables en termes de corrélation symbolique entre les valeurs de la variable à expliquer observées et prédites pour les données de calibration, mais sont plutôt en faveur de la *Symbolic Covariance Regression* pour les données de validation. Il faut noter que ces corrélations symboliques tiennent compte des deux bornes de l'intervalle conformément à l'Eq. (4). Les erreurs de prédiction sont données relativement aux prédictions des bornes inférieures et supérieures des intervalles associés à la variable à expliquer CONDEMN, selon l'Eq. (13) pour la méthode SCR. Il s'ensuit que ces erreurs différencient la méthode des centres comme étant la moins performante et celle présentant le plus de variabilité en comparaison aux deux autres méthodes de régression symbolique, en termes de calibration mais aussi de validation. Ces résultats sont comparables à ceux donnés par Lima Neto [23] pour la comparaison des méthodes des centres et des centres et des rangs. Par ailleurs, les méthodes des centres et des rangs (CRM) et *Symbolic Covariance Regression* ont des performances comparables, avec une légère faveur pour la méthode des centres et des rangs en termes de calibration et pour la *Symbolic Covariance Regression* en termes de validation. On peut noter, notamment pour l'échantillon de calibration, que l'erreur de prédiction de la méthode SCR présente une distribution plus dissymétrique (*i.e.*, valeurs parfois élevées de RMSE) que celle de la méthode CRM.

Par la suite, seuls les résultats de la *Symbolic Covariance Regression* sont interprétés. Les hypothèses d'application de la régression linéaire relatives aux résidus sont tout d'abord évaluées. Le test de Shapiro indique une hypothèse de normalité de la distribution des résidus acceptable pour les prédictions des bornes inférieures et supérieures (p-value$_{inf}$=0.70 ; p-value$_{sup}$=0.66), confirmée par les *qqplots* donnés Figure 4(a). Les tests de homoscédasticité et d'auto-corrélation des résidus n'ont pas encore développés pour les régressions symboliques, cependant l'évaluation visuelle des liens entre les valeurs de la variable à expliquer prédite et les résidus standardisés donnés Figure 4(b) montre une hétéroscédasticité acceptable, ainsi qu'une indépendance discutable pour les résidus de la borne inférieure de l'intervalle et acceptable ceux de la borne supérieure.

Les coefficients du modèle régression symbolique de la *Symbolic Covariance Regression*,

calculés sur l'ensemble des ($N_c = 15$) concepts, sont donnés par l'Eq. (14).

$$\begin{aligned}\textsc{Condemn} = 1.85 - 0.0028.\textsc{Mort} + 0.00027.\textsc{Area} - 0.025.\textsc{Freqchicken} \\ -0.0026.\textsc{Tlairage} - 0.023.\textsc{StockingD} + \textsc{Residus}\end{aligned} \quad (14)$$

Il s'ensuit que le taux de saisie à l'abattoir est principalement expliqué par la fréquence de visite de l'éleveur durant l'élevage (Freqchicken) ainsi que par la densité des animaux dans les caisses de transport (StockingD).

4 Conclusion et perspectives

Cet article propose d'appliquer une démarche complète d'Analyse de Données Symboliques à des données d'épidémiologie vétérinaire à valeurs d'intervalles. Deux méthodes originales sont appliquées : la *Symbolic Covariance PCA* et la *Symbolic Covariance Regression*. Celles-ci présentent l'avantage d'être 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. Ces méthodes ont été programmées sur le logiciel `R`. Dans le cadre de l'ACP, nous avons proposé une représentation factorielle des variables actives et supplémentaires, ce qui n'avait pas été développé auparavant. Par ailleurs, nous avons présenté une comparaison, par validation croisée sur la base d'un exemple, de trois méthodes de régression de référence : les méthodes des centres, des centres et des rangs et la *Symbolic Covariance Regression*. Sur notre exemple, les deux dernières méthodes ont de bonnes et comparables performances. Cependant, la dernière méthode présente un réel avantage en termes d'interprétation du fait de l'unicité de ses coefficients de régression. Ces méthodes pour variables à valeurs d'intervalles ainsi que leurs applications montrent une bonne adéquation de celles-ci aux données d'épidémiologie vétérinaire lorsque que l'unité statistique est l'élevage (mesures répétées sur les animaux) ; pour le cas où l'unité statistique est l'animal et que l'on souhaite s'affranchir des différences entre élevages, les analyses multigroupes sont à appliquer.

Malgré l'ensemble des avantages pré-cités de ces méthodes, l'hypothèse de distribution uniforme sur laquelle sont basés les calculs peut ne pas être vérifiée pour certains jeux de données. Dans le cadre des données d'épidémiologie vétérinaire, l'extension de l'Analyse de Données Symboliques aux variables à valeurs d'histogrammes, *i.e.*, pondération des bornes des intervalles par des probabilités d'appartenance, est donc souhaitable. Lors de ces cinq dernières années, de nombreux travaux ont été menés dans le but de prédire une distribution et non un simple réel comme il était proposé dans le premier modèle de régression pour variables à valeurs d'histogrammes [5]. Ce sont notamment les travaux de [30] et [12] qui ont abouti tout récemment à ceux de [16] et [13]. Ces nouveaux modèles de régression proposent de transformer chaque histogramme en une fonction quantile et exploitent au mieux de nouvelles métriques comme la distance de Wasserstein. D'autres travaux sont en cours de publication comme les extensions des méthodes *Symbolic Covariance* ACP et *Symbolic Covariance Regression* aux variables à valeurs d'histogrammes sur la base des estimations des statistiques de base par vraisemblance [19]. Tout comme il a été proposé pour les variables à valeurs d'intervalles [8], le développement de modèles probabilistes est une ouverture possible.

Références

[1] Afonso, F. : Méthodes prédictives par extraction de règles en présence de données symboliques. Thèse de doctorat. Université Paris Dauphine (2005)

[2] Bertrand, P., Goupil, F. : Descriptive statistics for symbolic data. Analysis of symbolic data : exploratory methods for extracting statistical information from complex data. Eds. H.H. Bock, E. Diday. Springer-Verlag, Berlin, (2000) 103–124

[3] Billard, L., Diday, E. : Regression analysis for interval-valued data. Data analysis, classification and related methods. Eds. H.A.L. Kiers, J.-P. Rassoon, P.J.F. Groenen, and M. Schader. Springer-Verlag, Berlin, (2000) 369–374

[4] Billard, L., Diday, E. : Symbolic regression analysis. Classification, clustering, and data analysis. Recent advances and applications. Eds. K. Jajuga, A. Sokolowski, H.H. Bock, Springer, Berlin, (2002) 281–288

[5] Billard, L., Diday, E. : Symbolic data analysis : conceptual statistics and data mining. Wiley, New York (2006)

[6] Billard, L. : Dependencies and variation components of symbolic interval-valued data. Selected contributions in data analysis and classification. Springer-Verlag, Berlin, (2007) 3–13

[7] Billard, L. : Sample covariance functions for complex quantitative data. Processing, World Conferences International Association of Statistical Computing, Yokohama, Japan (2008)

[8] Brito, P., Duarte, A. P. : Modelling interval data with normal and skew-normal distributions, Journal of Applied Statistics **39** (2011) 3–20

[9] de Carvalho, F.A.T., Lima Neto, E.A., Tenorio, C.P. : A new method to fit a linear regression model for interval-valued data. Lecture Notes in Computer Science, Advances in Artifical Intelligence. Springer-Verlag, (2004) 295–306

[10] Cazes, P., Chourakria, A., Diday, E., Schektman, Y. : Extension de l'analyse en composantes principales à des données de type intervalle. Revue de Statistique Appliquée, **45** (1997) 5–24

[11] Chouakria, A., Cazes, P., and Diday, E. : Symbolic principal component analysis. In : Analysis of symbolic data : explanatory methods for extracting statistical information from complex data. Eds. H.H. Bock, E. Diday. Springer-Verlag, Berlin, (2000) 200–212

[12] Dias, S., Brito, P. : A new linear regression model for histogram-valued variables. In : 58th ISI world statistics congress, Dublin, Ireland. http ://isi2011.congressplanner.eu/pdfs/950662 (2011)

[13] Dias, S., Brito, P. : Linear regression model with histogram-valued variables. Statistical Analysis and Data Mining, **8** (2015) 75–113

[14] Gioia, F., Lauro, C. : Principal component analysis on interval data. Computational Statistics, **21** (2006) 343–363

[15] Irpino, A., Lauro, C., Verde, R. : Visualizing Symbolic Data by Closed Shapes. In : Between Data Science and Applied Data Analysis. Eds. M. Schader, W. Gaul, M. Vichi, Berlin : Springer, (2003) 244–251

[16] Irpino, A., Verde, R : Linear regression for numeric symbolic variables : an ordinary least squares approach based on Wasserstein Distance. Advances in Data Analysis and Classification, **9** (2015) 81–106

[17] Lauro, C., Palumbo, F. : Principal component analysis of interval data : a symbolic data analysis approach. Computational Statistics, **15** (2000) 73–87

[18] Le-Rademacher, J., Billard, L. : Symbolic covariance principal component analysis and visualization for interval-valued data. Journal of computational and graphical statistics, **21** (2012) 413–432

[19] Le-Rademacher, J., Billard, L. : Likelihood functions and some maximum likelihood estimators for symbolic data. Journal of Statistical Planning and Inference, **14** (2011) 1593–1602

[20] Le-Rademacher, J. : Principal component analysis for interval-valued and histogram-valued data and likelihood functions and some maximum likelihood estimators for symbolic data. Thèse de doctorat. Université d'Athens (Georgia, USA) (2008)

[21] Lauro, N. C., Verde, R., Irpino, A. : Principal component analysis of symbolic data described by intervals, in Symbolic Data Analysis and the SODAS Software, Eds. E. Diday, M. Noirhomme-Fraiture, Chichester : Wiley, (2008) 279–311

[22] Liang, K.Y., Zeger, S.L. : Longitudinal data analysis using generalized linear models. Biometrika, **73** (1986) 13–22.

[23] Lima Neto, E.A., de Carvalho F.A.T. : Centre and range method for fitting a linear regression model to symbolic interval data. Computational Statistics and Data Analysis, **52** (2008) 1500–1515

[24] Lima Neto, E.A., de Carvalho F.A.T. : Constrained linear regression models for symbolic interval-valued variables. Computational Statistics and Data Analysis, **54** (2010) 333–347

[25] Lupo C., Chauvin C., Balaine L., Petetin I., Peraste J., Colin P., Le Bouquin S. : Post mortem condemnation of processed broiler chickens in Western France, Vet. Rec. **162** (2008) 709–713

[26] Makosso-Kallyth : Analyse en composantes principales de variables symboliques de type histogramme. Thèse de doctorat. Université Paris Dauphine (2010)

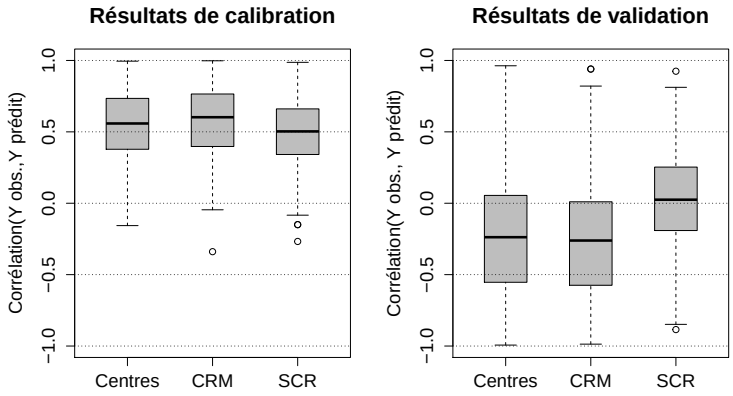
[27] Nelder, J.A., Wedderburn, R.W.M. : Generalized linear models, Journal of the Royal Statistical Society. Series A, **135** (1972) 370–384

[28] Palumbo, F., Lauro, C. : A PCA for interval-valued data based on midpoints and radii. In : New Developments in Psychometrics. Eds. H. Yanai, A. Okada, K. Shigemasu, Y. Kano, J. Meulman. Psychometric Society, Springer-Verlag, Tokyo, (2003) 641–648

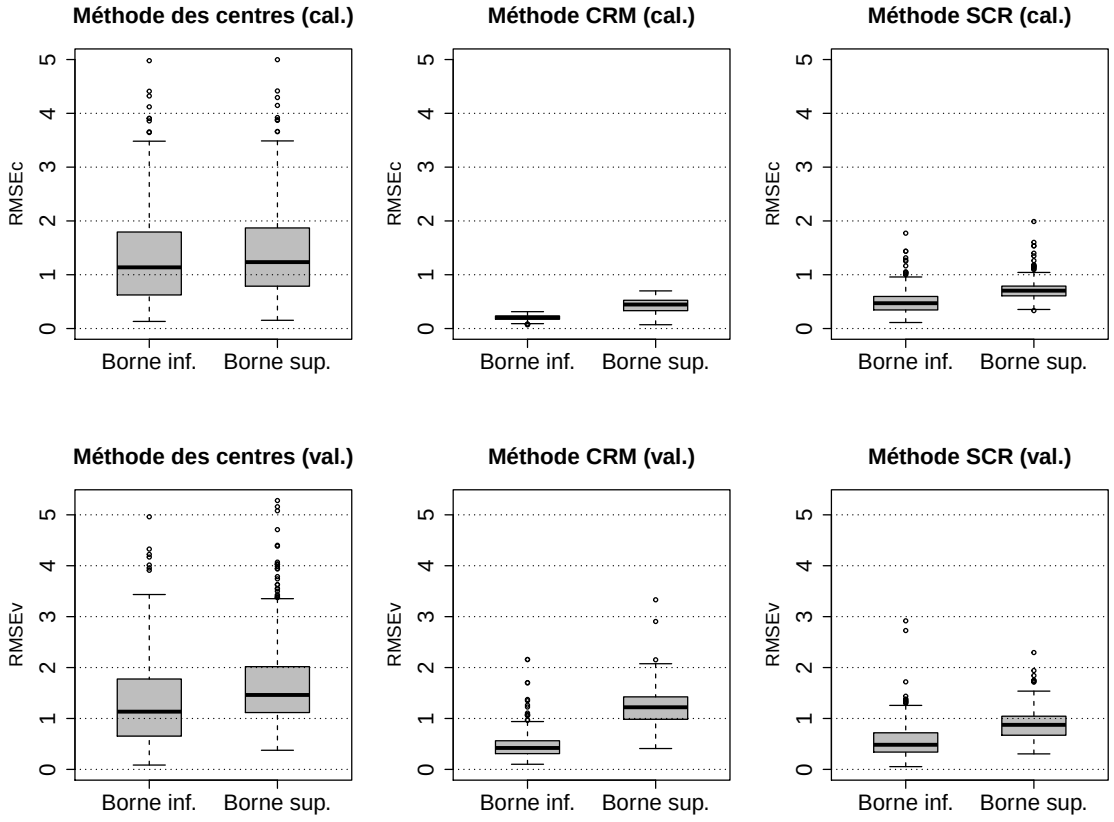
[29] Rodriguez, O.R., Calderon, O., Zuniga, R. : RSDA-R to Symbolic Data Analysis (1.2). URL http ://cran.r-project.org/web/packages/RSDA/. (2014)

[30] Verde, R., Irpino, A. : Ordinary least squares for histogram data based on Wasserstein distance. In : Proceedings of COMPSTAT'2010, Eds. Lechevallier, Y., Saporta, G., **60**, 581–588. Physica, Heidelberg (2010)

[31] Xu, W. : Symbolic data analysis : interval-valued data regression. Thèse de doctorat. Université d'Athens (Georgia, USA) (2010)

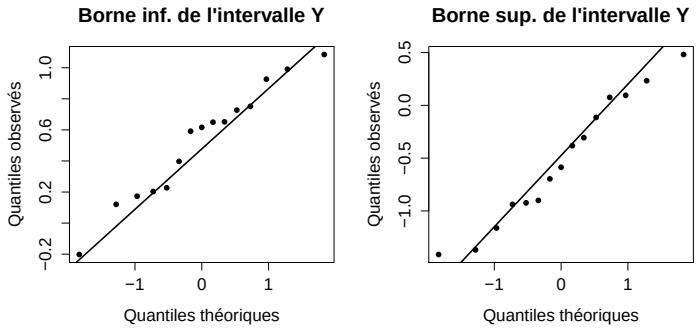


(a) Corrélation symbolique entre les valeurs de la variable à expliquer observée et prédite pour les données de calibration et de validation.

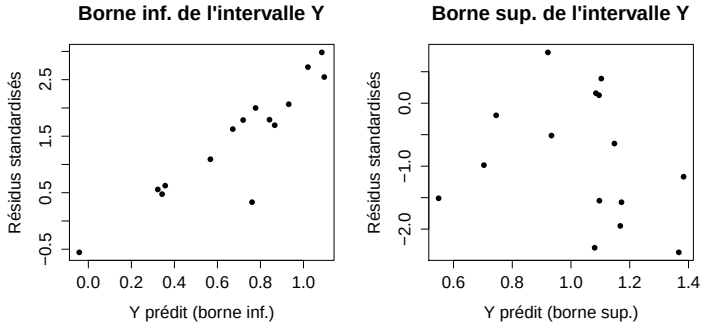


(b) Erreurs de calibration (cal.) et de validation (val.) relatives à la prédiction des bornes inférieures et supérieures de la variable à expliquer.

FIGURE 3 – Comparaison des performances de trois régressions symboliques pour variables à valeurs d'intervalles : la méthode des centres, la méthode des centres et des rangs (CRM) et la *Symbolic Covariance Regression* (SCR). Résultats issus d'une validation croisée basée sur 500 simulations. Données relatives aux saisies de poulets de chair à l'abattoir (N_c=15 concepts).



(a) Évaluation de la normalité des résidus des bornes inférieures et supérieures de la variable à expliquer (*qqplots*).



(b) Évaluation de l'auto-corrélation des résidus des bornes inférieures et supérieures de la variable à expliquer.

FIGURE 4 – Évaluation des hypothèses d'application de la régression linéaire relatives aux résidus de la *Symbolic Covariance Regression* (SCR). Données relatives aux saisies de poulets de chair à l'abattoir (N_c=15 concepts).