Estimation à noyau du tau de Kendall conditionnel pour l'analyse des interactions au sein des systèmes agricoles

Naomi Ouachene, Tristan Senga Kiessé, Michael S. Corson, Stéphane Girard

Estimation à noyau du tau de Kendall conditionnel pour l'analyse des interactions au sein des systèmes agricoles

Naomi Ouachene[1(a)], Tristan Senga Kiessé[1(b)], Michael S. Corson[1(c)] & Stéphane Girard[2]

[1] *UMR SAS, INRAE, Institut Agro, 35000 Rennes, France*
[(a)] *naomi.ouachene@inrae.fr* [(b)] *tristan.senga-kiesse@inrae.fr* [(c)] *michael.corson@inrae.fr*
[2] *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France, stephane.girard@inria.fr*

Résumé. La formalisation des relations entre de multiples variables est un enjeu majeur dans la modélisation des systèmes complexes tels que les systèmes agricoles. Les dépendances et interactions multiples au sein des systèmes d'élevage rendent difficiles la modélisation de l'effet des pratiques sur leur productivité et leurs performances environnementales. Cette étude a donc pour objectif de contribuer à une meilleure formalisation des relations entre un groupe de variables ciblées au sein de ces systèmes, en évaluant la corrélation entre deux variables conditionnellement aux valeurs prises par une troisième variable. Pour cela, nous avons étudié un estimateur du tau de Kendall conditionnel basé sur la méthode de lissage par noyau, qui intègre l'approche statistique des copules. Une application a été faite sur des données de systèmes d'élevage bovins laitiers collectées sur toute la France. On a notamment cherché à évaluer comment l'efficacité des pratiques de gestion déployées afin d'améliorer les performances environnementales de ces systèmes peut être affectée par l'existence d'interactions avec d'autres facteurs. La fenêtre de lissage a été sélectionnée par validation croisée et les effets de bord induits par le noyau utilisé ont été considérés. L'estimateur du tau de Kendall a montré une variation de la corrélation entre variables descriptives des systèmes d'élevage, d'où la nécessité de désigner des leviers permettant d'améliorer les performances environnementales adaptés au contexte et à la stratégie de gestion des exploitations.

Mots-clés. Système d'élevage, Ferme laitière, Performance environnementale, Interaction, Copule, Tau de Kendall conditionnel

Abstract. Formalizing relations among multiple variables is an important issue when modeling complex systems such as agricultural systems. Modeling management practices that influence farm performances is difficult since multiple dependencies and interactions occur in agricultural systems. This study aimed to improve formalization of multiple relations among a target group of descriptive variables of dairy farms by assessing correlation between two variables conditionally on a third one. Thus, we study a conditional Kendall's tau coefficient based on a kernel method that integrated the copula approach. We analyze data on French dairy cattle systems, aiming to assess how the effectiveness of management practices may vary due to interactions with other factors. The kernel bandwidth was selected using cross-validation, and edge effects caused by the type of kernel were considered. Estimating conditional Kendall's tau demonstrates how correlation among descriptive variables of a farming system may vary. Thus, it is necessary to highlight practices that can improve environmental performances that are adapted to a farm's context and management strategy.

Keywords. Farming system, Dairy farm, Environmental performance, Interaction, Copula, Conditional Kendall's tau

1 Introdution

Les systèmes d'élevage animaux sont à l'origine de 14,5 à 18 % des émissions de gaz à effet de serre (GES), dont les 3/4 sont imputées aux ruminants (Dumont et al. 2016). Par exemple, dans une ferme laitière, la combustion des carburants utilisés par les engins agricoles émet du dioxyde de carbone (CO_2). De même, la fermentation entérique ainsi que la gestion des effluents émettent du méthane (CH_4), tandis que la fertilisation azotée des cultures ou des prairies est la première source du protoxyde d'azote. Plusieurs actions sont susceptibles d'améliorer les performances environnementales des systèmes d'élevage bovins laitiers, généralement en modifiant les pratiques de gestion (Pellerin et al. 2013). Par exemple, une manière de réduire les émissions de CH_4 est de changer les rations des animaux, en diminuant la quantité de protéine ingérée que les vaches ne sont pas capables d'assimiler. Néanmoins, les interactions avec d'autres facteurs (par exemple les conditions météorologiques, ou d'autres pratiques de gestion) peuvent influencer l'efficacité des actions mises en place. Cela soulève la problématique de l'identification adéquate de ces interactions.

Des modèles mécanistes complexes ont été développés afin de représenter le fonctionnement d'un système d'élevage laitier dans son intégralité (Chardon 2008). Toutefois, il peut être nécessaire d'étudier les relations entre un groupe restreint de variables, pour mieux comprendre un phénomène local à l'intérieur du système. Par exemple, des mesures statistiques comme le tau de Kendall ou le rho de Spearman permettent d'évaluer la corrélation entre deux variables. Cependant, la corrélation entre deux variables peut augmenter ou diminuer conditionnellement à une covariable. Dans ce contexte, nous étudions un estimateur du taux de Kendall conditionnel récemment proposé par Derumigny et Fermanian (2019a). Cet estimateur se base sur un modèle de copules conditionnelles représentant la structure de dépendance entre deux variables conditionnellement à une troisième variable. Il utilise les poids non paramétriques de Nadaraya-Watson, qui sont basés sur une fonction noyau et un paramètre de lissage. Nous montrons que le potentiel des actions mises en place afin d'améliorer les performances environnementales des exploitations dépend grandement de la stratégie de gestion des fermes.

2 La méthode des copules

Nous présentons la méthode des copules pour modéliser la structure de dépendance entre deux variables aléatoires (v.a.) continues X_1 et X_2. On note par F_1 et F_2 les fonctions de répartition (f.d.r.) de X_1 et X_2, f_1 et f_2 les densités de probabilité associées ainsi que $F_{1,2}$ la f.d.r. jointe du couple (X_1, X_2). D'après le théorème de Sklar (1959), il existe alors une unique copule $C_{1,2} : [0,1]^2 \mapsto [0,1]$ telle que :

$$F_{1,2}(x_1, x_2) = C_{1,2}(F_1(x_1), F_2(x_2)),$$

ou encore,

$$C_{1,2}(u_1, u_2) = F_{1,2}(F_1^{-1}(u_1), F_2^{-1}(u_2)), \tag{1}$$

avec $u_1 = F_1(x_1)$ et $u_2 = F_2(x_2)$. De là, on comprend l'usage du terme "copule" qui vient du Latin "copula", en lien avec son action de "coupler" les lois marginales afin d'obtenir une f.d.r. jointe (Nelsen 2006). Dans le cas bivarié, une copule C respecte les propriétés suivantes (Nelsen 2006) :

(i) $C_{1,2}(u_1, 0) = C_{1,2}(0, u_2) = 0$,

(ii) $C_{1,2}(u_1, 1) = u_1$ et $C_{1,2}(1, u_2) = u_2$,

(iii) $C_{1,2}$ est 2-croissante, c'est-à-dire que pour $u_1, u_2, v_1, v_2 \in [0, 1]$ tels que $u_1 \leq u_2$ et $v_1 \leq v_2$, $C_{1,2}(u_2, v_2) - C_{1,2}(u_2, v_1) - C_{1,2}(u_1, v_2) + C_{1,2}(u_1, v_1) \geq 0$.

Les copules présentent de nombreux avantages. Par exemple, un modèle de copule $C_{1,2}$ est indépendant du choix des distributions marginales (Genest et Favre 2007). De plus, il existe différentes familles de copules pouvant modéliser une grande diversité de structures de dépendance, qu'elles soient non linéaires, ou spécifiquement entre les valeurs extrêmes (Nelsen 2006). Un modèle de copule $C_{1,2}$ qui lie de manière unique X_1 et X_2 permet également d'exprimer des mesures de corrélation comme le tau de Kendall (Genest et Favre 2007) :

$$\tau_{1,2} = 4 \int_{[0,1]^2} C_{1,2}(u_1, u_2) dC_{1,2}(u_1, u_2) - 1. \tag{2}$$

Le tau de Kendall se définit aussi comme la différence entre la probabilité d'obtenir une paire concordante et la probabilité d'obtenir une paire discordante, qui peut s'exprimer de manière empirique par :

$$g(\mathbf{X}_1, \mathbf{X}_2) = \mathbb{1}\{(X_{1,1} - X_{1,2})(X_{2,1} - X_{2,2}) > 0\} - \mathbb{1}\{(X_{1,1} - X_{1,2})(X_{2,1} - X_{2,2}) < 0\}, \tag{3}$$

avec $\mathbf{X}_i := (X_{1,i}, X_{2,i})_{i=\{1,2\}}$ deux versions indépendantes du couple $\mathbf{X} := (X_1, X_2)$.

3 Tau de Kendall conditionnel

Considérons deux v.a. continues X_1 et X_2, et Z une covariable continue. Les f.d.r. marginales de X_1 et X_2 conditionnellement à Z sont notées $F_{1|Z=z}(x_1) = \mathbb{P}(X_1 \leq x_1 | Z = z)$ et $F_{2|Z=z}(x_2) = \mathbb{P}(X_2 \leq x_2 | Z = z)$. De même, la f.d.r. jointe du couple (X_1, X_2) conditionnellement à Z est notée $F_{1,2|Z=z}(x_1, x_2) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2 | Z = z)$. Le théorème de Sklar (équation (1)) peut être appliqué pour définir une unique copule $C_{1,2|Z=z}$ telle que :

$$C_{1,2|Z=z}(u_1, u_2) = F_{1,2|Z=z}(F_{1|Z=z}^{-1}(u_1), F_{2|Z=z}^{-1}(u_2)), (u_1, u_2) \in [0, 1]^2. \tag{4}$$

Comme montré par Derumigny et Fermanian (2019a), le tau de Kendall conditionnel de X_1 et X_2 sachant Z se définit alors à partir des équations (2) et (4) :

$$\tau_{1,2|Z=z} = 4 \int_{[0,1]^2} C_{1,2|Z=z}(u_1, u_2) dC_{1,2|Z=z}(u_1, u_2) - 1.$$

Il existe différentes méthodes pour estimer le tau de Kendall conditionnel, telles que le lissage par noyau, celles basées sur les algorithmes de classification (Derumigny et Fermanian 2019a, 2019b), ou sur les modèles additifs généralisés (Vatter et Chavez-Demoulin 2015). Nous étudions un estimateur basé sur la méthode non paramétrique de lissage par noyau qui a été proposée par Derumigny et Fermanian (2019a) en se basant notamment sur les travaux de Gijbels et al. (2011) tel que :

$$\begin{aligned}\hat{\tau}_{1,2|Z=z} &= 4\int_{[0,1]^2} \hat{C}_{1,2|Z=z}(u_1,u_2)d\hat{C}_{1,2|Z=z}(u_1,u_2) - 1 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_{n,i}(z,h)w_{n,j}(z,h)\mathbb{1}\{x_{1,i} \leq x_{1,j}, x_{2,i} \leq x_{2,j}\} - 1, \end{aligned} \tag{5}$$

où $\hat{C}_{1,2|Z=z}$ est un estimateur de la copule conditionnelle, $h = h(n) > 0$ est la fenêtre de lissage telle que $\lim_{n\to+\infty} h(n) = 0$, et $w_{n,i}(\cdot, h)$ est la séquence de poids de Nadaraya-Watson telle que :

$$w_{n,i}(z,h) = \frac{K((Z_i - z)/h)}{\sum_{j=1}^{n} K((Z_j - z)/h)}.$$

Le noyau $K : \mathbb{R} \to \mathbb{R}$ est traditionnellement une densité de probabilité symétrique ($\int K(t)dt = 1$ et $K(-t) = K(t)$), de moyenne nulle ($\int tK(t)dt = 0$), de variance finie ($\int t^2K(t)dt < +\infty$) et de carré intégrable ($\int K^2(t)dt < +\infty$) (Tsybakov 2006). Dans les applications nous avons utilisé le noyau d'Epanechnikov.

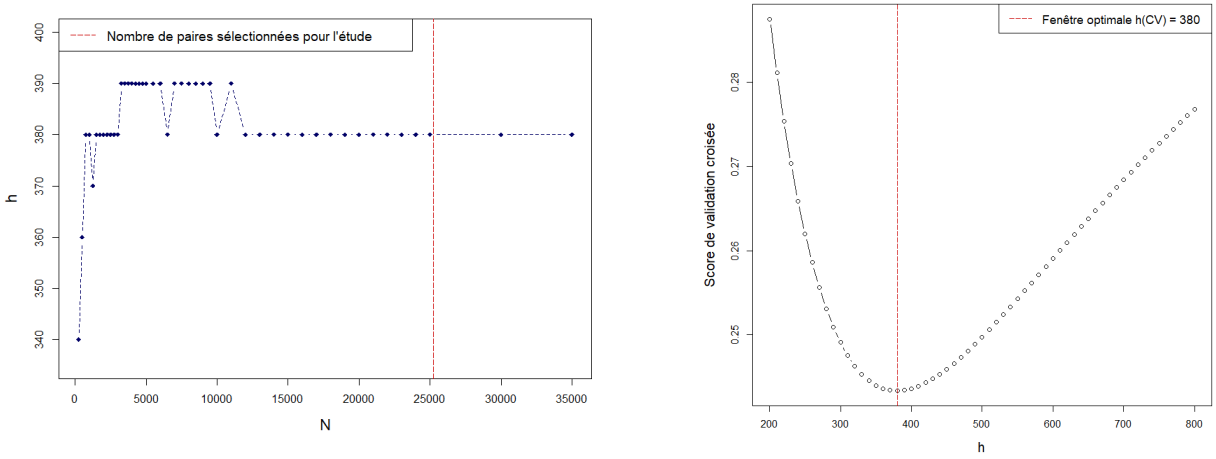
Le choix de la fenêtre de lissage à noyau h peut être vu comme un problème d'optimisation biais/variance. Une fenêtre choisie trop étroite va donner un tau de Kendall conditionnel trop erratique, tandis qu'une fenêtre choisie trop grande va gommer les évolutions du tau conditionnel, et donner une courbe trop lisse. Le choix de la fenêtre optimale h_{CV} se fait par validation croisée suivant la méthode "leave-one-out" (Derumigny et Fermanian 2019a). Le score de la validation croisée, minimisé par la fenêtre h_{CV} optimale, se définit comme suit :

$$CV(h) = \frac{1}{N}\sum_{k=1}^{N}\left[g(\mathbf{X}_{i_k}, \mathbf{X}_{j_k}) - \hat{\tau}_{1,2;-(i_k,j_k)|Z=(Z_{i_k}+Z_{j_k})/2}\right]^2,$$

où

- les N paires $(\mathbf{X}_i, \mathbf{X}_j)_k$ sont construites pour $i, j \in \{1, \ldots, n\}$ et $k \in \{1, \ldots, N\}$,
- $\hat{\tau}_{1,2;-(i_k,j_k)|Z=(Z_{i_k}+Z_{j_k})/2}$ représente une estimation du tau de Kendall conditionnel du couple (X_1, X_2) auquel on a retiré les valeurs de la paire k sélectionnée, sachant la moyenne $Z = (Z_i + Z_j)/2$ des valeurs de la covariable Z associées à cette paire.

Ainsi, le score de validation croisée se définit comme la différence entre le tau de Kendall empirique de chaque paire (équation (3)), et une estimation du tau de Kendall conditionnel (équation (5)) sur le reste du jeu de données. Pour ne pas tenir compte des effets de bord, nous n'avons considéré que les taux conditionnels estimés sur la plage $[\min(Z) + h_{CV}, \max(Z) - h_{CV}]$. La méthode décrite ci-dessus est disponible sous `R` dans la bibliothèque `CondCopulas` (Derumigny 2022).



(a) (b)

FIGURE 1 – Evolution de la fenêtre optimale de lissage par noyau en fonction du nombre N de paires choisies (a) et sélection par validation croisée de la fenêtre h_{CV} pour un N fixé (b).

4 Application et résultats

Les données utilisées dans cette étude sont issues d'une enquête réalisée auprès de 2523 fermes en France en 2013, dans le cadre du projet LIFE Carbon Dairy conduit par l'Institut de l'Elevage (IDELE). Le premier couple de variables étudié est la production laitière brute par vache (l/vache/an) et les émissions de CH_4 entérique par unité de gros bétail [1] (UGB) (t éq.CO_2/UGB/an). Ce couple est corrélé de manière significative, avec un tau de Kendall de 0,69. Ainsi, une grande production de lait implique une quantité plus importante de CH_4 entérique émis, en partie dû à l'alimentation des animaux. La covariable considérée est la quantité de matière organique digestible (kg/vache/an) contenue dans la ration des animaux, puisqu'elle constitue un levier potentiel pour réduire le CH_4 entérique issu de la production animale (Pellerin et al. 2013). Cette covariable est corrélée significativement avec la production laitière ($\tau = 0,89$) et les émissions de CH_4 entérique ($\tau = 0,75$).

L'approche de validation croisée mise en oeuvre a nécessité le choix adéquat du nombre de paires à sélectionner, afin d'obtenir une fenêtre h optimale et stable (Figure 1a). Dans notre cas, la valeur par défaut du nombre de paires était de $10 \times n$, qui nous a conduit à la sélection d'une fenêtre $h_{CV} = 380$ (Figure 1b). La corrélation conditionnelle estimée entre la production laitière et les émissions de CH_4 entérique croît ($\hat{\tau}_{1,2|Z=z} \in [-0,57; 0,20]$) entre 2408 et 3062 kg de quantité de matière organique contenue dans la ration annuelle de chaque vache, elle passe de 0,20 à 0,12 entre 3062 et 3352 kg/vache/an, avant de se stabiliser autour de 0,40 pour une quantité de matière organique digestible supérieure à 4000 kg/vache/an (Figure 2). De plus, le tau de Kendall conditionnel estimé est toujours inférieur au tau de Kendall non conditionnel du couple. Les taux conditionnels estimés aux bords de la covariables sont

1. Unité de référence utilisée pour agréger différentes espèces de bétail, d'âges et de besoins nutritionnels variés.

illustrés (en gris) dans la Figure 2.

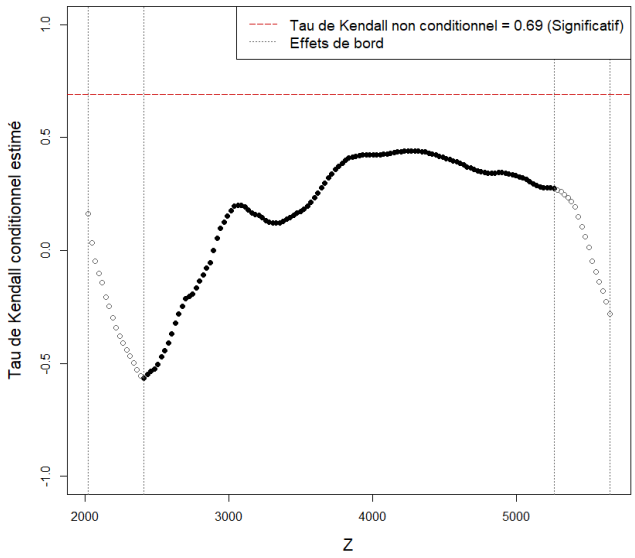


FIGURE 2 – Estimation du tau de Kendall conditionnel de la production laitière par vache et des émissions de CH_4 entérique par UGB conditionnellement à la quantité en matière organique digestible de la ration, avec représentation (points en gris) des tau conditionnels situés aux bords.

Les variations des tau de Kendall conditionnels estimés discriminent les exploitations dont les stratégies de gestion diffèrent. L'efficacité des leviers mis en place afin d'améliorer les performances environnementales va ainsi varier en fonction du contexte de chaque exploitation. Des analyses devront être approfondies pour mieux caractériser les spécificités des fermes, en intégrant notamment leurs conditions météorologiques. D'un point de vue plus technique, il serait intéressant de tester des noyaux asymétriques continus tels que Beta ou Gamma (Chen 1999, 2000) dans l'estimation du tau de Kendall conditionnel pour réduire les effets de bord. De même, d'autres méthodes de choix de fenêtre peuvent être étudiées telles que la "règle-du-pouce", ou encore la validation croisée par "k-folds" (Derumigny et Fermanian 2019a).

Bibliographie

Chardon, X. (2008), Evaluation environnementale des exploitations laitières par modélisation dynamique de leur fonctionnement et des flux de matière : Développement et application du simulateur MELODIE *Thèse de doct.*, Agro Paris Tech.

Cheng, X.S. (1999), Beta kernel estimators for density functions, *Computational Statistics & Data Analysis*, 31.2, pp. 131-145.

Cheng, X.S. (2000), Probability Density Function Estimation Using Gamma Kernels, *Annals of the Institute of Statistical Mathematics*, 52, pp. 471-480.

Derumigny, A. (2022), CondCopulas : Estimation and Inference for Conditional Copula Models, *R package version 0.1.2.*, `https://CRAN.R-project.org/package=CondCopulas`.

Derumigny, A. et Fermanian, J.-D. (2019a), On kernel-based estimation of conditional Kendall's tau : finite-bounds and asymptotic behavior, *Dependence Modeling*, 7.1, pp. 292-321.

Derumigny, A. et Fermanian, J.-D. (2019b), A classification point-of-view about conditional Kendall's tau, *Computational Statistics & Data Analysis*, 135, pp. 70-94.

Dumont, B. et al. (2016), Rôles, impacts et services issus des élevages en Europe, *Synthèse de l'expertise scientifique collective (INRA)*.

Genest, C. et Favre A.-C. (2007), Everything You Always Wanted to Know about Copula Modeling but Were Afraid to Ask, *Journal of Hydrologic Engineering*, 12.4, pp. 347-368.

Gijbels, I. et al. (2011), Conditional copulas, association measures and their applications, *Computational Statistics & Data Analysis*, 55.5, pp. 1919-1932.

Nelsen B., R. (2006), An Introduction to Copulas, 2^e *éd. Springer Series in Statistics*.

Pellerin, S. et al. (2013), Quelle contribution de l'agriculture française à la réduction des émissions de gaz à effet de serre ? Potentiel d'atténuation et coût de dix actions techniques, *Synthèse (INRA)*.

Sklar, A. (1959), Fonctions de répartition à n dimensions et leurs marges, *Publication de l'Institut de Statistique de L'Université de Paris, 8*, pp. 229-231.

Tsybakov, A. B. (2004), Introduction à l'estimation non paramétrique, 1^{re} *éd. Springer Berlin, Heidelberg*.

Vatter, T. et Chavez-Demoulin, V. (2015), Generalized additive models for conditional dependences structures, *Journal of Multivariate Analysis*, 141, pp. 147-167.