Explorer l'influence conjointe de prédicteurs fonctionnels sur une réponse réelle via une régression pénalisée

Girault Gnanguenon Guesse, Patrice Loisel, Bénedicte Fontez, Thierry Simonneau, Nadine Hilgert

Explorer l'influence conjointe de prédicteurs fonctionnels sur une réponse réelle via une régression pénalisée

Girault Gnanguenon Guesse [1], Patrice Loisel [2], Bénedicte Fontez [3], Thierry Simonneau [4] & Nadine Hilgert [5]

[1] *MISTEA, Université Montpellier, Institut Agro, INRAE, Montpellier, France.*
girault-bogues.gnanguenon-guesse@inrae.fr
[2] *MISTEA, Université Montpellier, Institut Agro, INRAE, Montpellier, France.*
patrice.loisel@inrae.fr
[3] *MISTEA, Université Montpellier, Institut Agro, INRAE, Montpellier, France.*
benedicte.fontez@supagro.fr
[4] *LEPSE, Université Montpellier, Institut Agro, INRAE, Montpellier, France.*
thierry.simonneau@inrae.fr
[5] *MISTEA, Université Montpellier, Institut Agro, INRAE, Montpellier, France.*
nadine.hilgert@inrae.fr

Résumé. En agronomie, l'avènement de nouveaux capteurs permet d'observer à haute fréquence des dynamiques de variables agro-environnementales affectant la production. Cette nouvelle situation nécessite de faire appel à d'autres outils statistiques ou de les révolutionner afin de tirer de la connaissance de ces données dites fonctionnelles. Dans un contexte où la production est affectée par un effet combiné complexe des différentes dynamiques de variables agro-environnementales, nous proposons une nouvelle approche exploitant des distributions conjointes de variables fonctionnelles pour expliquer une variable réelle (scalaire) représentant un facteur de production. Les simulations effectuées permettent de mettre en exergue une approche exploratoire se rapprochant des techniques de type boosting qui permet d'identifier diverses distributions conjointes associées aux courbes explicatives, d'y associer des coefficients via des régressions linéaires pénalisées et structurées puis de retenir une distribution conjointe optimale expliquant au mieux la variable à prédire. Cette approche a aussi l'avantage de pouvoir intégrer au besoin dans la modélisation des connaissances dites "connaissances d'experts" provenant de la littérature ou autres afin d'améliorer la fiabilité de l'approche statistique proposée. Cette approche qui se veut exploratoire peut être utilisée comme un modèle prédictif sous certaines conditions. Développée à la base pour l' agronomie, cette approche est générique et peut être utilisée pour résoudre des problèmes de type scalar-on function avec comme hypothèse principale l'identification d'effets combinés de variables explicatives fonctionnelles. Une limite de cette approche est un risque de surestimation mais divers critères permettent d'y pallier. L'utilisation de l'approche pour analyser des données réelles permet d'identifier des combinaisons de classes de température - irradiance et de moments de la journée affectant l'accumulation d'anthocyanes et de polyphénols dans la baie de raisin.

Mots-clés. exploration de données fonctionnelles, distribution conjointe, régression linéaire pénalisée, critères d'information, agronomie.

Abstract. In agronomy, the development of new sensors has allowed to observe at high frequency the dynamics of agri-environmental variables affecting production. This new situation

requires using other statistical tools or revolutionizing them in order to learn from this so-called functional data. In a context where production is affected by a complex combined effect of these different dynamics of agri-environmental variables, we propose a new approach using joint distributions of functional variables to explain a real (scalar) variable representing a production factor. Simulations carried out highlight an exploratory approach closed to boosting techniques that identifies various joint distributions associated to the explanatory curves, associates coefficients to each of them via penalized and structured linear regressions and then selects an optimal joint distribution that best explains the variable to be predicted. This approach has the additional advantage of being able to integrate, if necessary, so-called "expert knowledge" from the literature or other sources into the modeling process in order to improve the reliability of the proposed statistical approach or advise on these "expert knowledge". This exploratory approach can be used as a predictive model under certain conditions. Developed initially for agronomy, it is generic and can be used to solve scalar-on-function problems with the main goal of identifying combined effects of functional explanatory variables. One limitation of this approach is the risk of overestimation, but various criteria are available inside the approach to overcome it. Using the approach to analyze real data allows to identify associations of temperature - irradiance and time that affect the accumulation of anthocyanins and polyphenols in grape berries.

Keywords. functional data mining, joint distribution, penalized linear regression, information criteria, agronomy

1 Introduction

De nos jours, plusieurs domaines d'activités sont révolutionnés par l'avènement des données massives. Ces données massives sont considérées de diverses manières parmi lesquelles la grande famille des données fonctionnelles regroupant courbes, spectres, images, etc. Selon Ferraty et Vieu (2006), une variable aléatoire $\mathcal{X}$ est dite fonctionnelle si ses valeurs sont dans un espace de dimension infinie et une observation X de $\mathcal{X}$ est appelée donnée fonctionnelle. En réalité, seulement quelques points discrets du phénomène continu sont observés $\mathcal{X} = \{X(t) \ : \ t \in T\}$.

L'un des axes majeurs de recherche autour de ces données concerne leur implication dans des problèmes de régression. Dans la littérature, ces problèmes sont habituellement classés en 3 catégories en fonction du rôle joué par les données fonctionnelles (Reiss et *al.* (2010) ; Ramsay et Silverman (2005)). On distingue les régressions "scalar-on-function", "function-on scalar" et "function-on-function". Dans cet exposé, nous nous intéresserons au problème de type 'scalar-on-function' où la variable à prédire est un scalaire et le prédicteur, une fonction. Plus précisément, nous nous intéresserons à un problème où les prédicteurs peuvent être deux ou plusieurs variables fonctionnelles. Diverses méthodes existent pour résoudre les régressions de type 'scalar-on-function' et le lecteur pourra se référer à Reiss et *al.* (2017) qui en présente une revue. L'approche proposée nommée SPICEFP (Sparse and Structured Procedure to Identify Combined Effects of Functional Predictors) permet de s'intéresser spécifiquement à l'hypothèse d'influence conjointe des prédicteurs fonctionnels. Nous la présentons brièvement dans la suite de ce texte et présenterons également quelques résultats.

2 L'approche proposée

Considérons deux variables explicatives fonctionnelles que sont $\tau = \{\tau_i(t) : t \in T;\ i = 1, \ldots, n$ et $\mathcal{I} = \{\mathcal{I}_i(t) : t \in T;\ i = 1, \ldots, n\}$, toutes deux des fonctions observées aux mêmes instants t. Considérons d'un autre côté, une variable réponse $y_i \in \mathbb{R},\ i = 1, \ldots, n$ que l'on souhaiterait expliquer par τ et $\mathcal{I}$ en faisant l'hypothèse d'une influence conjointe des deux variables explicatives fonctionnelles sur y. L'approche SPICEFP permet d'atteindre cet objectif. Sa mise en oeuvre nécessite cinq étapes à savoir :

1. transformer (catégoriser) des variables fonctionnelles :
   - *la catégorisation* : pour un individu i fixé, catégorisons en n_τ ($n_\mathcal{I}$) classes l'observation τ_i ($\mathcal{I}_i$) suivant une échelle linaire. Les $n_\tau + 1$ ($n_\mathcal{I} + 1$) bornes de classes nécessaires sont : $l(v),\ v = 1, 2, \ldots, n_\tau + 1$ ($L(w),\ w = 1, 2, \ldots, n_\mathcal{I} + 1$). Leurs valeurs peuvent être obtenues par l'équation (2.1). Les modalités utilisés pour la catégorisation de τ_i ($\mathcal{I}_i$) s'écrivent sous la forme $[l(v), l(v+1)[, v = 1, \ldots, n_\tau$ ($[L(w), L(w+1)[, w = 1, \ldots, n_\mathcal{I}$).
   $$l(v) = \underline{\tau} + (v-1)\left(\frac{\bar{\tau} - \underline{\tau}}{n_\tau}\right),\ v = 1, \ldots, n_\tau + 1 \tag{2.1}$$
   avec $\underline{\tau} \in \mathbb{R}$ et $\bar{\tau} \in \mathbb{R}$ les valeurs (réelles) minimale et maximale de τ. Précisons que n_τ ($n_\mathcal{I}$) est à fixer afin de calculer $l(v)$ ($L(w)$).
   - *l'obtention d'une distribution conjointe en effectif* : à partir d'un tableau de contingence C_i^u, de dimension $(n_\tau \times n_\mathcal{I})$ dont les valeurs $C^u_{i,(v,w)}$ sont obtenues via (2.2). Les modalités de la distribution conjointe en effectif obtenues sont notées $[l(v), l(v+1)[\_[L(w), L(w+1)[, v = 1, \ldots, n_\tau,\ w = 1, \ldots, n_\mathcal{I}$. Elles seront appelées modalités conjointes et sont au nombre de $n_\tau \times n_\mathcal{I}$.
   $$C^u_{i,(v,w)} = \sum_{t=1}^{T} \mathbb{1}_{\tau_i(t) \in [l(v), l(v+1)[,\ \mathcal{I}_i(t) \in [L(w), L(w+1)[}\ ;\ v = 1, \ldots, n_\tau\ ;\ w = 1, \ldots, n_\mathcal{I} \tag{2.2}$$
   avec $C^u_{i,(v,w)} \in \mathbb{N}$; $C_i^u \in \mathbb{N}^{n_\tau \times n_\mathcal{I}}$; $(\tau_i(t), \mathcal{I}_i(t)) \in \mathbb{R}^2$; $\sum_{v=1}^{n_\tau} \sum_{w=1}^{n_\mathcal{I}} C^u_{i,(v,w)} = Card(T)$ et $u = (n_\tau, n_\mathcal{I}) \in \mathbb{N}^2$
   - *la vectorisation (empilement colonne après colonne) et la transposition du tableau de contingence* C_i^u : elles donnent
   $$X_i^u = {}^\intercal Vect(C_i^u);\ X_i^u \in \mathbb{R}^{(n_\tau n_\mathcal{I})} \tag{2.3}$$
   un vecteur ligne de longueur $n_\tau \times n_\mathcal{I}$ qui représente pour u fixé, le nombre d'instants t au cours desquels un individu i a été observé dans chacune des $n_\tau \times n_\mathcal{I}$ conditions décrites par les modalités conjointes. On obtient ainsi une matrice X^u, dont chaque ligne X_i^u correspond à un individu. A cette matrice de nouvelles variables explicatives X^u, on rajoute l'information relative à la proximité des modalités conjointes crées en utilisant un graphe $G^u(V^u, E^u)$ où V^u représente X^u et E^u l'ensemble des arêtes liant deux modalités conjointes proches. Deux modalités conjointes sont dites proches si les classes suivant la variable τ (indexées par v) ou [1] les classes suivant la variable $\mathcal{I}$

1. ou exclusif

(indexées par w) sont consécutives comme le montre la figure 1. Cette figure présente un exemple de catégorisation d'un couple de variables longitudinales en $n_\tau \times n_\mathcal{I} = 3 \times 3$ modalités conjointes ainsi qu'une une identification par 4 flèches des modalités conjointes voisines de $[v_2, v_3[\_[w_2, w_3[$ indexées par $j : (v = 2, w = 2)$

- *la construction de diverses distributions conjointes en vue d'en identifier une optimale* : ne connaissant pas a priori la valeur optimale du vecteur u que nous noterons u^*, il est proposé de l'identifier dans le cadre de la mise en oeuvre de l'approche. Étant donné qu'à partir d'un vecteur de catégorisation u nous obtenons une matrice X^u et un graphe $G^u(V^u, E^u)$ associé, explorer différents vecteurs u revient donc à explorer différents graphes explicatifs, tous issus des deux variables fonctionnelles τ et $\mathcal{I}$.

2. effectuer un Generalized Fused Lasso (Tibshirani et Taylor (2011)) sur chaque graphe indexé par u à partir de l'équation (2.4) :

$$\beta^{u,GFL} = \underset{\beta^u}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{n} (y_i - X_i^u \beta^u)^2 \right\} + \lambda_p^u \sum_{v=1}^{n_\tau} \sum_{w=1}^{n_\mathcal{I}} |\beta^u| + \lambda_f^u \sum_{(j,k) \in E^u} |\beta_j^u - \beta_k^u| \quad (2.4)$$

avec : $\lambda_p^u \geq 0$ et $\lambda_f^u \geq 0$ les paramètres de régularisation à optimiser. Pour $j = (v, w)$ fixé, les couples (j, k) relatifs à j et contenus dans E sont $(j, k)_1 = ((v, w), (v + 1, w))$ et $(j, k)_2 = ((v, w), (v, w + 1))$.

3. utiliser un critère pour choisir le meilleur graphe de prédicteurs $G^{u^*}(V^{u^*}, E^{u^*})$ et les coefficients estimés $\hat{\beta^{u^*}}$ associés
4. calculer les résidus associés au meilleur modèle $\varepsilon^{u^*} = y_i - X_i^{u^*} \hat{\beta^{u^*}}$
5. vérifier les conditions d'arrêts pour :
   - retourner à l'étape 2 en remplaçant la variable à prédire par les résidus du meilleur modèle ε^{u^*} obtenus à l'étape 4
   - ou arrêter l'approche

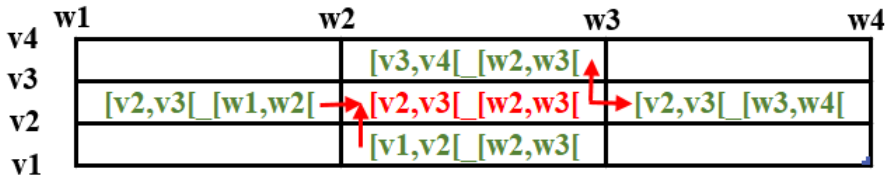

FIGURE 1 – Voisinage pris en compte dans le cadre du Generalized Fused LASSO

3 Quelques résultats

Nous avons tout d'abord illustré notre approche avec des simulations pour bien comprendre ses caractéristiques. Nous nous sommes donnés pour cela des variables explicatives de température et d'irradiance (issues d'expérimentations du projet européen INNOVINE, pour étudier

les effets combinés d'une exposition des baies de raisin plus ou moins forte au soleil et d'une température de l'air normale ou élevée de quelques degrés.) ainsi qu'un vecteur de coefficients parcimonieux. La variable à prédire (variable fictive pour les simulations) a été simulée en associant le tableau de contingence des variables explicatives, le vecteur de coefficients qu'il s'agissait d'estimer et un bruit Gaussien. La figure 2 illustre à gauche un exemple de vecteur de coefficients simulé. Dans cet exemple, on simule une influence positive au coeur d'une zone d'influence négative pour des valeurs faibles de température et d'irradiance sur la variable à prédire. Les cases blanches correspondent à aucune influence (coefficients nuls) des variables explicatives et les cases noires correspondent aux modalités conjointes n'ayant jamais été observées dans le cadre de ces données. L'objectif ici est d'identifier un modèle qui respecte la parcimonie simulée et estime les zones d'influence négative et positive, tout en tenant compte de la continuité des valeurs d'une case à l'autre. L'estimation des coefficients avec notre approche est illustrée avec le graphique de droite. L'estimation des zones d'influence et non influence sont relativement bien reproduites.

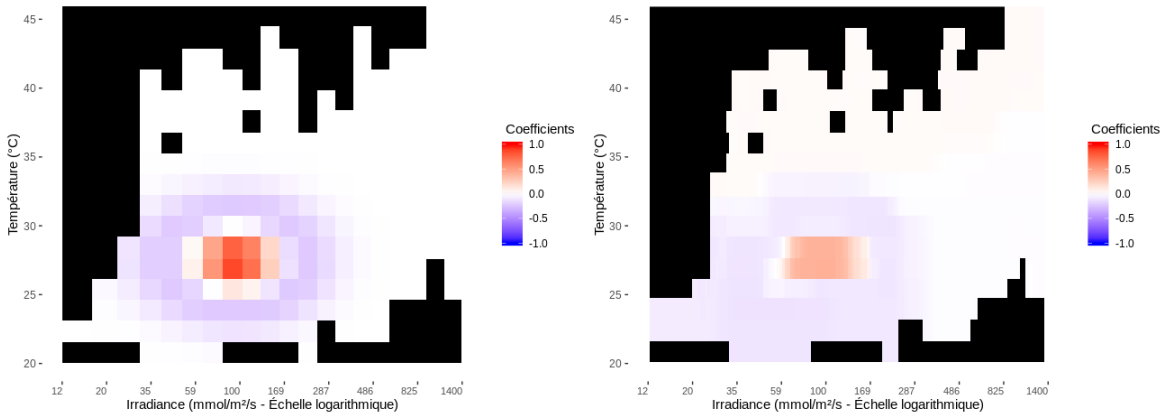

FIGURE 2 – Coefficients simulés (à gauche) et coefficients estimés (à droite).

Cette approche a été appliquée sur les variations hebdomadaires d'indices de Ferari (variables réelles issues des expérimentations d'INNOVINE, qui sont une mesure non destructive de la qualité des baies de raisin). Les résultats seront présentés lors de la conférence.

4 Conclusion

SPICEFP est une approche exploratoire utilisant des outils de la statistique inférentielle. Son but primordial est de fournir des modalités conjointes participant à l'explication d'une variable réponse. Elle sous-entend une influence conjointe des prédicteurs et est conçue pour une exploration dans ce sens. L'un de ses atouts est sa capacité à identifier à une nouvelle itération, une nouvelle distribution conjointe permettant de mieux expliquer la variable à prédire. Ce faisant, on augmente les risques de surestimation. D'où une sensibilité de l'approche à la surestimation lorsque l'erreur de mesure associée à la variable à prédire est élevée. Deux techniques développées lors de la conception de l'approche permettent d'identifier les cas de surestimation.

5 Remerciements

Ce travail a bénéficié d'une aide de l'État gérée par l'Agence Nationale de la Recherche au titre du programme d'Investissements d'Avenir portant la référence ANR-16-CONV-0004 (DigitAg). Les données présentées ont été acquises au cours du projet INNOVINE, financé par le septième programme-cadre de la Communauté européenne (FP7/2007-2013), dans le cadre de la convention de subvention No. FP7-311775.

Bibliographie

Ferraty, F. et Vieu, P. (2006), *Nonparametric Functional Data Analysis : Theory and Practice*, Springer Series in Statistics, Springer-Verlag, New York.

Ramsay J. et Silverman B.W. (2005), *Functional Data Analysis*, Springer Series in Statistics, Springer.

Reiss, P.T., Goldsmith J., Shang H.L. et Ogden R.T. (2017). Methods for scalar-on-function regression, *International Statistical Review*, 85(2), pp. 228–249.

Reiss, P.T., Huang L. et Mennes M. (2010). Fast function-on-scalar regression with penalized basis expansions, *The international journal of biostatistics*, 6(1), article 28.

Tibshirani, R.J. et Taylor, J. (2011). The solution path of the generalized lasso, *Ann. Statist.*, 39(3), pp. 1335-1371.