# Sélection d'intervalles pour des prédicteurs fonctionnels à partir de forêts aléatoires

Rémi Servien, Nathalie Vialaneix

▶ **To cite this version:**

Rémi Servien, Nathalie Vialaneix. Sélection d'intervalles pour des prédicteurs fonctionnels à partir de forêts aléatoires. Journées de Statistique de la SFdS, Jun 2022, Lyon, France. hal-03697845

**HAL Id: hal-03697845**
**https://hal.science/hal-03697845**

Submitted on 17 Jun 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Sélection d'intervalles pour des prédicteurs fonctionnels à partir de forêts aléatoires

Rémi Servien [1,2] & Nathalie Vialaneix [3]

[1] *INRAE, Univ. Montpellier, LBE, 102 Avenue des étangs, F-11000 Narbonne, France*
[2] *ChemHouse Research Group, Montpellier, France*
[3] *Université de Toulouse, INRAE, UR MIAT, F-31320, Castanet-Tolosan, France*

**Résumé.** Nous nous intéressons ici au problème de sélection de variables dans un cadre de régression fonctionnelle. Le but est de sélectionner des points de mesure consécutifs afin de déterminer les intervalles importants dans la prédiction de la variable cible. Pour cela nous nous basons sur les forêts aléatoires et évaluons des variantes possibles pour trois étapes de l'approche générale que nous proposons (création de groupes, définition des résumés, sélection) que nous comparons sur des données simulées et réelles.

**Mots-clés.** Régression fonctionnelle, sélection d'intervalles, forêts aléatoires.

**Abstract.** In this communication, we focus on the problem of variable selection in a functional regression framework. Our goal is to select consecutive measurement points to define the most important intervals for prediction in functional regression. The proposed approach is based on random forest and we evaluate different variants for three steps defining our approach (interval creation, summary definition, selection), which are compared on both simulated and real datasets.

**Keywords.** Functional regression, interval selection, random forest.

## 1 Contexte

Dans de nombreuses applications, les données se présentent sous la forme de vecteurs de très grande dimension qui sont la discrétisation d'un phénomène continu pour lesquelles des mesures ont été réalisées : données météorologiques (« courbes » de températures et précipitations), spectres en chimiométrie haut-débit, ... Ces données sont généralement désignées sous le terme commun de *données fonctionnelles* et, depuis longtemps, tout un pan de la littérature statistique s'intéresse à adapter les analyses statistiques aux spécificités de ce type de données [Ramsay and Silverman, 1997, Ferraty and Vieu, 2006].

Dans cette proposition de communication, nous nous intéresserons, en particulier, à un problème de sélection de variables adapté au cadre fonctionnel. Ce travail se place dans le contexte de la régression fonctionnelle, dans lequel une variable réelle, $Y$, doit être prédite par une variable aléatoire fonctionnelle, $X$. Les approches de sélections multi-dimensionnelles de sélection de variables peuvent être utilisées sur la discrétisation observée de la variable $X$ mais donnent alors des résultats peu interprétables et qui ne

rendent pas compte de la nature spécifique des données. En effet, cette approche tend à produire des sélections de points de mesure de $X$ isolés alors que la variable cible $Y$ est intrinsèquement dépendante d'un ou plusieurs intervalles constitués de points de mesure consécutifs.

Nous proposons ici une approche basée sur la méthode des forêts aléatoires [Breiman, 2001] qui a montré ses bonnes performances dans les problèmes de prédiction en grande dimension. Nous proposons une approche en trois étapes et évaluons des variantes possibles pour ces trois étapes. L'intégralité de la procédure est testée sur un problème simulé (utilisant des courbes de données météorologiques réelles) et sur un problème réel dans lequel la vérité terrain est connue.

# 2 Description de l'approche proposée

Dans la suite, on notera $(X, Y)$, le couple de variables aléatoires dans lequel $X$ est une covariable fonctionnelle et $Y$ est une variable catégorielle ou continue à prédire. On dispose de $n$ observations i.i.d. de $(X, Y)$, $(x_i, y_i)_{i=1,\dots,n}$ pour lequelles $x_i$ est observée partiellement aux temps de mesure $t_1$, $t_2$, …, $t_d$. On note alors $\mathbf{x}_i = (x_i(t_1), \dots, x_i(t_d))^\top \in \mathbb{R}^d$.

L'approche proposée se définit en trois étapes :

— la **création de groupes** de prédicteurs ;
— la **définition de variables résumées** pour chacun de ces groupes ;
— la **sélection de variables**.

## 2.1 Création de groupes de prédicteurs

L'objectif de cette étape est de fournir aux modèles prédictifs une partition de $\{t_1, \dots, t_d\}$ en groupes de temps de mesures adjacents. La difficulté réside ici dans le fait que, pour $d$ temps de mesure, le nombre de partitions possibles des prédicteurs ($2^d$) est rapidement impossible à explorer de manière exhaustive. Aussi, nous proposons d'utiliser des approches gloutonnes permettant d'explorer une hiérarchie de regroupements.

Parmi les approches sélectionnées, explicitées dans l'algorithme 1 , nous testons :

— la classification ascendante hiérarchique sous contrainte de contiguïté qui utilise la perte d'inertie intra-groupes minimale basée sur la corrélation entre temps de mesure, $\mathrm{Cor}((x_i(t_j))_i^\top, (x_i(t_{j+1}))^\top)$ comme implémentée dans **adjclust** [Ambroise et al., 2019] ;
— une version sous contrainte de contiguïté de l'approche **ClustOfVar** [Chavent et al., 2012] ;
— une approche basée sur la régression linéaire dans laquelle l'étape 1 est basée sur le critère

$$\arg \min_{j=1,\dots,p-T} \mathrm{MSE}(X(t_j), X(t_{j+1}) - \max\left(\mathrm{MSE}(X(t_j)),\ \mathrm{MSE}(X(t_{j+1}))\right.$$

où MSE(.) est la valeur du critère des moindres carrés pour la prédiction de $Y$ à partir de, respectivement $(X(t_j), X(t_{j+1})$, $X(t_j)$ et $X(t_{j+1})$.

**Algorithm 1** Création de groupes par approche gloutonne

$\mathcal{D}^0 = (C_i^0)_{1 \leq i \leq p}$ with $C_i^0 = \{x_i\}$ ▷ Initialisation
**for** $T = 1$ to $p - 1$ **do**
    Trouver le temps de mesure $t_{j^\star}$ ($j^\star \in \{1, \ldots, p - T\}$) tel que le groupe de temps de mesure $\{t_{j^\star}, t_{j^\star+1}\}$ est optimal pour un certain critère ▷ Meilleur candidat
    **for** $j = 1$ to $p - T - 1$ **do** ▷ Mise à jour de $\mathcal{D}^{T-1}$ en $\mathcal{D}^T$
        **if** $j < j^\star$ **then** $C_j^T = C_j^{T-1}$
        **else if** $j = j^*$ **then** $C_j^T = C_j^{T-1} \cup C_{j+1}^{T-1}$
        **else if** $j > j^\star$ **then** $C_j^T = C_{j+1}^{T-1}$
        **end if**
    **end for**
**end for**

Notons que les deux premières approches définissent des groupes de points d'observations de $X$ complètement non supervisés alors que la dernière utilise la capacité prédictive des points d'observations pour définir les groupes. Cette étape fournit donc $D$ partitions de prédicteurs, $\mathcal{D}_1, \ldots, \mathcal{D}_D$, où, dans notre cas, $D$ est toujours égal à $d$.

## 2.2 Définition de variables résumées

Pour chaque partition $\mathcal{D}_k$ ainsi obtenue, la **définition de variables résumées** consiste à résumer les variables $X(t_j)_{j \in C_\ell}$ pour un $C_\ell \in \mathcal{D}_k$. Pour chacune des partitions issues de chacune des trois méthodes présentées plus haut, nous définissons trois approches permettant de définir un résumé :

— comme dans [Deng et al., 2013] ("Time Series Forest"), chaque groupe est résumé par la valeur moyenne et l'écart type sur les temps de mesure inclus dans le groupe (deux variables créés par groupe, approche appelée « basique » par la suite) :

$$\tilde{X}_\ell^1 = \frac{1}{|C_\ell|} \sum_{j \in C_\ell} X(t_j) \qquad \text{et} \qquad \tilde{X}_\ell^2 = \sqrt{\frac{1}{|C_\ell|} \sum_{j \in C_\ell} (X(t_j) - \tilde{X}_\ell^1)^2} \; ;$$

— comme dans [Poterie et al., 2019, Menze et al., 2011], chaque groupe est résumé en utilisant une information utilisant la variable cible $Y$ : meilleure combinaison linéaire au sens de la régression linéaire, ridge, ou PLS ;

— lorsque les groupes ont été définis avec la version contrainte de **ClustOfVar**, comme [Chavent et al., 2021], il est aussi possible de résumer les points de mesure du groupe en utilisant la variable synthétique du groupe [Chavent et al., 2012].

### 2.3 Sélection de variables

Pour sélectionner les intervalles importants, nous utilisons deux approches de sélection de variables adaptées aux forêts aléatoires : l'approche **VSURF** [Genuer et al., 2010] (essentiellement basé sur l'utilisation de la notion d'importance) et l'approche **Boruta** [Kursa and Rudnicki, 2010], basée sur une approche de knockoffs [Barber and Candès, 2015]. Ces approches de sélection sont confrontées à une approche utilisant directement l'importance des variables pour repérer les intervalles importants.

## 3 Simulations

### 3.1 Données

L'intégralité des comparaisons est effectuée sur deux problèmes :

— **Données simulées** : nous avons utilisé les mêmes séries climatiques que décrites dans [Picheny et al., 2019] ($n = 1000$ séries temporelles mesurées à $t \in [\![287, 730]\!]$). Ces données ont été générées à l'aide du simulateur de données météorologiques WACSGen [Flecher et al., 2010], reproduisant le climat de Lleida, Catalogne, Espagne entre 1981 et 1982, pour générer les données selon le modèle (inspiré par [Grollemund et al., 2019]) :

$$\forall\, i = 1, \dots, 1000, \qquad y_i = \log\left(1 + |\langle x_i, \beta \rangle|\right) + \epsilon_i \tag{1}$$

où $\beta(t) = 4 \times \mathbf{1}_{\{t \in [320,410]\}} + 2 \times \mathbf{1}_{\{t \in [500,550]\}} - \mathbf{1}_{\{t \in [680,730]\}}$ et $\epsilon_i \sim \mathcal{N}(0, 0.5)$ sont i.i.d. ;

— **Truffes** : ce jeu de données est décrit en détails dans [Baragatti et al., 2019]. Il consiste en 25 années de production de truffes (rendement en kilos) pour 4 variables climatiques associées (température maximale, précipitation, bilan hydrique et déficit hydrique) fournies sous la forme de séries temporelles et mesurées mensuellement du mois de janvier de l'année $n$ au mois de mars de l'année $n + 1$. Une connaissance experte sur ces données permet de savoir quelles périodes doivent être retrouvées comme importantes, celles-ci étant potentiellement différentes pour chaque variable explicative.

### 3.2 Quelques résultats et discussions

Nous comparons les différentes combinaisons des étapes 1 et 2 par un test de différence des valeurs d'importance entre les points de mesure connus pour être importants et ceux connus pour ne pas l'être (test de Wilcoxon). Le tableau 1 donne la $p$-valeur minimale (sur l'ensemble de la hiérarchie de groups telle que définie dans la section 2.1) pour chaque combinaison de méthodes (ainsi que le nombre de groupes associés) pour une des variables des données simulées.

| Création | Définition | $p$-valeur ($-\log_{10}$) | Nombre de groupes |
|---|---|---|---|
| adjclust | basique | -48.3 | 24 |
| | modèle linéaire | 46.1 | 43 |
| | PLS | 34.7 | 64 |
| | ridge | 40.6 | 69 |
| ClustOfVar | basique | -45.9 | 101 |
| | ClustOfVar | 51.8 | 109 |
| | modèle linéaire | 44.4 | 118 |
| | PLS | 43.0 | 112 |
| | ridge | 46.4 | 118 |
| modèle linéaire | basique | -12.0 | 369 |
| | modèle linéaire | 17.2 | 52 |
| | PLS | 17.2 | 52 |
| | ridge | 14.8 | 55 |

TABLE 1 – **Données simulées.** Résultats comparatifs en terme de $p$-valeurs (voir le texte) et de nombre de groupes..

Les $p$-valeurs sont très petites, ce qui est encourageant. La meilleure méthode semble être l'enchaînement des deux étapes avec ClustOfVar. Des résultats similaires ont été obtenus pour toutes les variables des données « truffes » : la méthode ClustOfVar est systématiquement la meilleure pour créer les groupes. Si le nombre de groupes semblent, en revanche, important, il faut noter qu'il ne correspond pas complètement aux nombres d'intervalles qui seront finalement considérés comme « importants » (plusieurs groupes « importants » pouvant être adjacents et ne constituer qu'un seul intervalle « important »). La figure 1 montre les valeurs d'importance sur tout l'ensemble de définition, pour l'approche ClustOfVar, soit en utilisant l'importance initiale, soit ne conservant que les valeurs d'importance supérieure à $|\min_{C^t \in \mathcal{D}^t} \mathcal{I}_{\mathcal{C}^t}|$ où $\mathcal{I}_c$ est l'importance du résumé correspondant au groupe $c$ (approche dite « seuillée »). Celles-ci sont comparées à la vraie variable $\beta$ telle que définie dans l'équation (1).

Ces résultats sont prometteurs car nous retrouvons globalement les bons intervalles : les deux premiers intervalles, en partculier, sont presque parfaitement retrouvés mais le dernier est, par contre, perdu. Nous espérons que l'étape de sélection, que nous n'avons pas encore évaluée à l'heure d'écriture de cette présentation, permettra d'améliorer les résultats qui seront présentés de manière complète lors de la communication orale.

# Remerciements

Nous remercions la plateforme bioinformatique Genotoul Bioinfo `http://bioinfo.genotoul.fr/` pour la mise à disposition des ressources de calcul. Nous remercions également le métaprogramme INRAE DIGIT-BIO pour le financement du projet PhenoDyn qui soutient ces recherches. Enfin, nous remercions Pierre Casadebaig, Ronan Trepos, Victor Picheny, Meili Baragatti et François Le Tacon pour la mise à disposition des données de climats simulées et des données « truffes ».

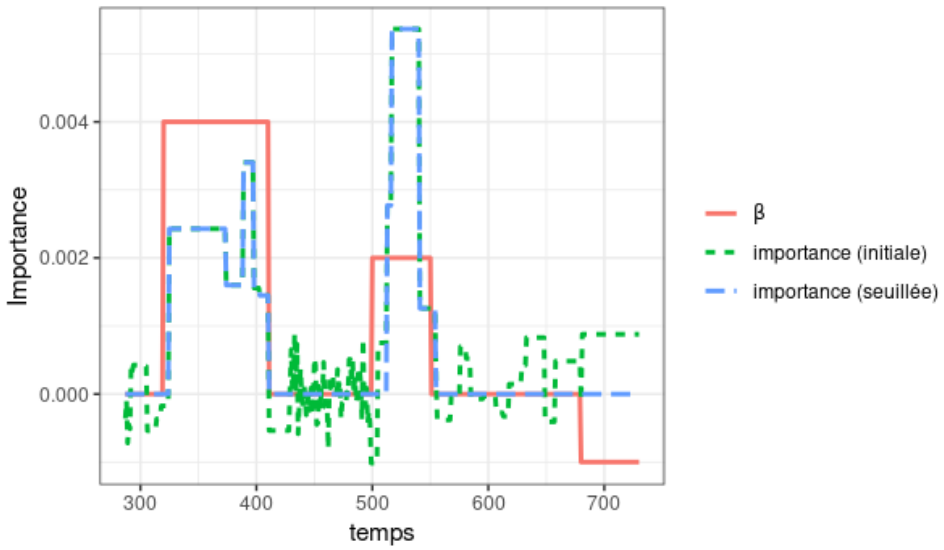


FIGURE 1 – **Données simulées.** Importances initiale et seuillée, comparées à la valeur de $\beta$, prédicteur ayant permis de générer les données.

# Bibliographie


[Ambroise et al., 2019] Ambroise, C. et al. (2019). *Algorithm Mol Biol*, 14 :22. doi.

[Baragatti et al., 2019] Baragatti, M. et al. (2019). *Mycorrhiza*, 29(2) :113–125. doi.

[Barber and Candès, 2015] Barber, R. and Candès, E. (2015). *Ann Stat*, 43(5) :2055–2085. doi.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Mach Learn*, 45(1) :5–32. doi.

[Chavent et al., 2012] Chavent, M. et al. (2012). *J Stat Softw*, 50(13) :1–16. doi.

[Chavent et al., 2021] Chavent, M. et al. (2021). *Commun Stat Simul Comput*, 50(2) :426–445. doi.

[Deng et al., 2013] Deng, H. et al. (2013). *Inf Sci*, 239 :142–153. doi.

[Ferraty and Vieu, 2006] Ferraty, F. and Vieu, P. (2006). *NonParametric Functional Data Analysis*. Springer. doi.

[Flecher et al., 2010] Flecher, C. et al. (2010). *Wat Resour Res*, 46(7) :W07519. doi.

[Genuer et al., 2010] Genuer, R. et al. (2010). *Patt Recognit Lett*, 31(14) :2225–2236. doi.

[Grollemund et al., 2019] Grollemund, P.-M. et al. (2019). *Bay Anal*, 14(1) :111–135. doi.

[Kursa and Rudnicki, 2010] Kursa, M. and Rudnicki, W. (2010). *J Stat Softw*, 36(11) :1–13. doi.

[Menze et al., 2011] Menze, B. H. et al. (2011). In *Proc of Mach Learn and Know Disc*, volume 6912, pages 453–469. Springer Berlin Heidelberg. doi.

[Picheny et al., 2019] Picheny, V. et al. (2019). *Stat Comput*, 29(2) :255–267. doi.

[Poterie et al., 2019] Poterie, A. et al. (2019). *Comp Stat*, 34 :1613–1648. doi.

[Ramsay and Silverman, 1997] Ramsay, J. and Silverman, B. (1997). *Functional Data Analysis*. Springer Verlag, New York. doi.