# Sélection de variables groupées avec les forêts aléatoires. Application à l'analyse des données fonctionnelles multivariées.

Baptiste Gregorutti[12], Bertrand Michel[2] & Philippe Saint Pierre[2]

[1] *Safety Line*
*15 rue Jean-Baptiste Berlier, 75013 Paris, France*

[2] *LSTA, Université Pierre et Marie Curie - Paris VI*
*Boîte 158, Tour 15-25, 2ème étage*
*4 place Jussieu, 75252 Paris Cedex 05, France*
*baptiste.gregorutti@safety-line.fr*

**Résumé.** Dans cet article, nous étudions la sélection de groupes de variables avec les forêts aléatoires. Dans un premier temps, nous introduisons une nouvelle mesure d'importance pour des groupes de variables. Nous étudions théoriquement cette mesure pour un modèle de régression additive. Nous montrons en particulier qu'en toute généralité, l'importance d'un groupe ne peut s'écrire comme la somme des importances individuelles des variables le composant. Dans une seconde partie, nous présentons une approche originale de sélection de variables en analyse de données fonctionnelles. En particulier, lorsque l'on observe un grand nombre de covariables à valeurs dans un espace de fonctions, chacune de ces variables peut être vue comme le groupe formé par ses coefficients de base (ondelettes, ACP fonctionnelle, etc.). Nous proposons donc d'utiliser l'importance groupée et un algorithme pas-à-pas pour sélectionner les covariables fonctionnelles. Cette méthode est appliquée au problème de l'analyse des données des enregistreurs de vol pour la prédiction des risques opérationnels en aéronautique.

**Mots-clés.** Forêts aléatoires, Importance des variables, Sélection de variables groupées, Analyse des données fonctionnelles

**Abstract.** In this paper, we study the selection of grouped variables using the random forests algorithm. We first propose a new importance measure adapted for groups of variables. Theoretical insights of this criterion are given for additive regression models. The second contribution of this paper is an original method for selecting functional variables based on the grouped variable importance measure. When we observe a large number of functional variables, we propose to regroup all of the basis coefficients (wavelet, functional PCA, etc.) and use a wrapper selection algorithm with these groups. The method is applied to a real life problem coming from aviation safety.

**Keywords.** Random Forests, Variable Importance, Grouped variable Selection, Functional data analysis

# 1 Introduction

Dans un contexte d'apprentissage en grande dimension, toutes les variables explicatives ne sont pas nécessairement importantes pour la prédiction de la variable d'intérêt. En effet, les variables non informatives peuvent avoir un effet néfaste sur la précision du modèle. Les techniques de sélection de variables fournissent une réponse naturelle à ce problème en éliminant les covariables qui n'apportent pas assez d'informations prédictives au modèle. La réduction du nombre de variables explicatives présente un double avantage. D'une part, un modèle contenant peu de variables est plus interprétable. D'autre part, l'erreur de prédiction se trouve réduite de fait de la suppression de variables non informatives.

Dans notre article Gregorutti et al [2], nous nous plaçons dans le contexte où les variables explicatives sont structurées en groupes de variables. Ce contexte statistique est motivé par des considérations pratiques. En effet, le choix de regrouper certaines variables est conduit par l'a priori que le statisticien a sur les données. Par exemple, il peut être pertinent de regrouper des variables corrélées ou des variables ayant des caractéristiques physiques communes.

Si la sélection de variables groupées a largement été étudiée dans le cas du modèle linéaire, notamment la méthode *Group Lasso* (Yuan, Lin [4]), ce problème n'a pas été considéré dans le cas des forêts aléatoires (Breiman [1]). Cet algorithme non paramétrique présente d'excellentes performances en pratique et peut être utilisé dans un contexte de sélection de variables au moyen de mesures d'importance. Intégrées à des algorithmes pas-à-pas, les forêts aléatoires sont une alternative non linéaire aux méthodes de type Lasso.

Les contributions de ce travail sont doubles. Dans un premier temps, nous définissons une nouvelle mesure d'importance pour des groupes de variables basée sur le même principe que l'importance par permutation définie par Breiman [1]. Une étude théorique de ce critère permet de montrer qu'en toute généralité, l'importance d'un groupe ne peut s'écrire comme la somme des importances individuelles des variables le composant.

La seconde contribution de ce travail est une nouvelle approche de l'analyse des données fonctionnelles (Ramsay, Silverman [3]) où l'on observe plusieurs covariables à valeurs dans un espace de fonctions. Une démarche classique dans ce contexte est de projeter chaque variable fonctionnelle sur un sous-espace de dimension finie engendré par une base de fonctions. Par exemple les bases de splines, d'ondelettes où d'ACP fonctionnelles sont classiquement utilisées dans la littérature (voir par exemple Ramsay, Silverman [3]). Notre approche vise à utiliser la mesure d'importance groupée dans ce contexte. En effet, chaque variable fonctionnelle est vue comme un groupe formé par ses coefficients de base. Nous sommes donc en mesure de proposer une procédure de sélection de variables fonctionnelles.

## 2 Mesure d'importance groupée avec les forêts aléatoires

On considère une variable d'intérêt $Y$ à valeur dans $\mathbb{R}$ et un vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_p)$. La régression vise à estimer la fonction $f(x) = \mathbb{E}[Y|\mathbf{X} = x]$ à partir d'un échantillon d'apprentissage $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ de $n$ vecteurs aléatoires indépendants et de même loi que $(\mathbf{X}, Y)$ où $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$. L'erreur commise par un estimateur $\hat{f}$ est alors $R(\hat{f}) = \mathbb{E}\left[(\hat{f}(\mathbf{X}) - Y)^2\right]$. Cette quantité étant inconnue en pratique, nous en considérons un estimateur empirique basé un échantillon de validation $\bar{\mathcal{D}}$ :

$$\hat{R}(\hat{f}, \bar{\mathcal{D}}) = \frac{1}{|\bar{\mathcal{D}}|} \sum_{i:(\mathbf{X}_i, Y_i) \in \bar{\mathcal{D}}} (Y_i - \hat{f}(\mathbf{X}_i))^2.$$

Les forêts aléatoires sont une méthode non paramétrique très compétitive pour l'estimation de $f$. Introduites par Breiman en 2001, elles consistent en l'agrégation d'un grand nombre d'arbres aléatoires basés sur une partition dyadique et récursive de l'espace des observations, ici $\mathbb{R}^p$. Plus précisément, $M$ arbres aléatoires sont construits à partir d'échantillons bootstrap $\mathcal{D}_n^1, \ldots, \mathcal{D}_n^M$ des données d'apprentissage $\mathcal{D}_n$. En conséquence, une collection d'estimateurs $\hat{f}_1, \ldots, \hat{f}_M$ de $f$ est considérée.

Une différence majeure dans la construction des arbres constituant la forêt par rapport aux algorithmes initiaux est la suivante : le critère de découpe intervenant dans le partitionnement est optimisé sur un faible nombre de variables choisi aléatoirement. L'estimateur final de la forêt est alors défini comme la prédiction moyenne de chaque arbre ainsi randomisé. L'aléa induit par l'échantillonnage bootstrap ainsi que le choix aléatoire des variables à chaque étape du partitionnement permet à l'estimateur agrégé de la forêt d'être meilleur que les arbres pris individuellement.

L'algorithme des forêts aléatoires propose également des critères permettant d'évaluer l'importance des covariables sur la prédiction de $Y$. Nous considérons ici la mesure d'importance par permutation due à Breiman [1]. Une variable $X_j$ est considérée comme importante pour la prédiction de $Y$ si en brisant le lien entre $X_j$ et $Y$, l'erreur de prédiction augmente. Pour briser le lien entre $X_j$ et $Y$, Breiman propose de permuter aléatoirement les valeurs de $X_j$. Plus formellement, considérons une collection d'ensembles "out-of-bag" (OOB) $\{\bar{\mathcal{D}}_n^m = \mathcal{D}_n \setminus \mathcal{D}_n^m,\ m = 1, \ldots, M\}$ contenant les observations non retenues dans les échantillons bootstrap. Ces ensembles seront utilisés pour calculer l'erreur de chaque arbre $\hat{f}_m$ . À partir de ces ensembles, définissons les ensembles out-of-bag permutés $\{\bar{\mathcal{D}}_n^{mj},\ m = 1, \ldots, M\}$ en permutant les valeurs de la $j$-ème variable des échantillons out-of-bag. La mesure d'importance par permutation est alors définie par

$$\hat{\mathcal{I}}(X_j) = \frac{1}{M} \sum_{m=1}^{M} \left[ \hat{R}(\hat{f}_m, \bar{\mathcal{D}}_n^{mj}) - \hat{R}(\hat{f}_m, \bar{\mathcal{D}}_n^{m}) \right]. \tag{1}$$

Cette quantité est l'équivalent empirique de la mesure d'importance $\mathcal{I}(X_j)$ comme l'ont formulé récemment Zhu et al. [5] :

$$\mathcal{I}(X_j) = \mathbb{E}\left[\left(Y - f(\mathbf{X}_{(j)})\right)^2\right] - \mathbb{E}\left[(Y - f(\mathbf{X}))^2\right], \tag{2}$$

où $\mathbf{X}_{(j)} = (X_1, \ldots, X'_j, \ldots, X_p)$ est un vecteur aléatoire tel que $X'_j$ est une réplication indépendante de $X_j$. La permutation de $X_j$ dans la définition de $\hat{\mathcal{I}}(X_j)$ revient donc à remplacer $X_j$ par une variable indépendante et de même loi dans (2).

Dans ce travail, nous supposons que l'ensemble des variables $(X_1, \ldots, X_p)$ est structuré en $K$ groupes. Plus formellement, soit $J = (j_1, \ldots, j_k)$ un vecteur de $k$ indices de $\{1, \ldots, p\}$ et $\mathbf{X}_J = (X_{j_1}, \ldots, X_{j_k})$ le sous-vecteur de $\mathbf{X}$ associé. L'importance du groupe $\mathbf{X}_J$ est définie par

$$\hat{\mathcal{I}}(\mathbf{X}_J) = \frac{1}{M}\sum_{m=1}^{M}\left[\hat{R}(\hat{f}_m, \bar{\mathcal{D}}_n^{mJ}) - \hat{R}(\hat{f}_m, \bar{\mathcal{D}}_n^{m})\right].$$

À la différence du critère (1), l'ensemble $\bar{\mathcal{D}}_n^{mJ}$ représente l'ensemble OOB dans lequel les valeurs du groupe $J$ sont aléatoirement permutées. L'importance théorique est alors donnée par :

$$\mathcal{I}(\mathbf{X}_J) = \mathbb{E}\big[Y - f(\mathbf{X}_{(J)})\big]^2 - \mathbb{E}\big[Y - f(\mathbf{X})\big]^2, \tag{3}$$

où $\mathbf{X}_{(J)} = (X_1, \ldots, X'_{j_1}, X_{j_1+1}, \ldots, X'_{j_2}, X_{j_2+1}, \ldots, X'_{j_k}, X_{j_k+1}, \ldots, X_p)^\top$ est tel que $\mathbf{X}'_J = (X'_{j_1}, X'_{j_2}, \ldots, X'_{j_k})^\top$ est un réplication indépendante de $\mathbf{X}_J$.

Dans la suite de cette Section, nous donnons des éléments théoriques du critère (3) pour l'importance du groupe $\mathbf{X}_{(J)}$. Notons tout d'abord $\mathbf{X}_{\bar{J}}$, le vecteur des variables n'appartenant pas à $\mathbf{X}_{(J)}$. Supposons que la distribution du vecteur $(X, Y)$ satisfait le modèle de régression

$$\begin{aligned} Y &= f(\mathbf{X}) + \varepsilon \\ &= f_J(\mathbf{X}_J) + f_{\bar{J}}(\mathbf{X}_{\bar{J}}) + \varepsilon, \end{aligned} \tag{4}$$

où $f_J$ et $f_{\bar{J}}$ sont deux fonctions mesurables, et $\varepsilon$ est une variable aléatoire telle que $\mathbb{E}[\varepsilon|\mathbf{X}] = 0$ et $\mathbb{E}[\varepsilon^2|\mathbf{X}]$ est finie.

**Proposition 1.** *Sous le modèle* (4)*, l'importance du groupe $J$ s'écrit*

$$\mathcal{I}(\mathbf{X}_J) = 2\,\mathrm{Var}\left[f_J(\mathbf{X}_J)\right].$$

Le corolaire suivant donne l'importance groupée pour des modèles plus spécifiques.

**Corollary 1.** *Sous le modèle* (4)*,*

1. *Si* $f_J(\mathbf{x}_J) = \sum_{j\in J} f_j(x_j)$ *et si* $(X_j)_{j\in J}$ *sont indépendantes, alors*

$$\mathcal{I}(\mathbf{X}_J) = 2\sum_{j\in J} \mathrm{Var}\left(f_j(X_j)\right) = \sum_{j\in J} \mathcal{I}\left(X_j\right).$$

2. *Si* $f_J(\mathbf{x}_J) = \sum_{j\in J} \alpha_j x_j$, *alors*

$$\mathcal{I}(\mathbf{X}_J) = 2\alpha_J^\top \mathrm{Cov}(\mathbf{X}_J)\alpha_J,\ \alpha_J = (\alpha_j)_{j\in J}.$$

Ce résultat montre que sous l'hypothèse d'additivité de la fonction de régression et si les variables du groupe sont indépendantes, alors l'importance groupée se décompose en la somme des importances individuelles. Comme le montre le Point 2 du Corolaire, cette propriété est perdue dès que les variables sont corrélées . Dans Gregorutti et al [2], nous comparons numériquement l'importance groupée aux importances individuelles pour plusieurs modèles. Ces simulations suggèrent qu'en toute généralité, l'importance groupée peut difficilement s'écrire comme la somme des importances individuelles.

# 3 Application à l'analyse des données fonctionnelles multivariées

Dans cette Section, nous présentons une approche originale de l'analyse des données fonctionnelles basée sur la mesure d'importance groupée définie précédemment. En particulier, nous considérons le problème de la sélection de variables de $p$ covariables fonctionnelles pour la prédiction d'une variable aléatoire réelle $Y$.

Supposons observer $p$ variables aléatoires $X^1, \ldots, X^p$ à valeurs dans l'espace $L^2([0,1])$ muni du produit scalaire

$$\langle f, g\rangle_{L^2} = \int f(t)g(t)dt,$$

pour $f, g \in L^2([0,1])$.

Une approche naturelle consiste à projeter chaque variable fonctionnelle sur un sous-espace de $L^2([0,1])$ de dimension finie, c'est-à-dire

$$\begin{aligned} X^u(t) &= \sum_{k=1}^{\infty} \langle X^u, \varphi_k\rangle_{L^2}\varphi_k(t), \\ &\simeq \sum_{k=1}^{d_u} \langle X^u, \varphi_k\rangle_{L^2}\varphi_k(t), \end{aligned}$$

où $\{\varphi_1, \varphi_2, \ldots\}$ est une base de fonctions orthogonales. Il s'agit ensuite de considérer les coefficients de base $\langle X^u, \varphi_1\rangle_{L^2}, \ldots, \langle X^u, \varphi_{d_u}\rangle_{L^2}$ comme nouvelles variables explicatives

dans un algorithme d'apprentissage. Ces variables aléatoires réelles résument l'information portée par $X^u$ dans le sous-espace de $L^2([0,1])$ de dimension $d_u$. Dans Gregorutti et al [2], nous étudions le cas des bases d'ondelettes.

Dans ce contexte, sélectionner les variables fonctionnelles $X^1, \ldots, X^p$ revient à sélectionner les groupes formés par leurs coefficients de base. Nous proposons alors l'algorithme de sélection pas-à-pas suivant :

1. Projeter chaque variable fonctionnelle sur un sous-espace de dimension finie ;
2. Construire une forêt et calculer l'erreur ;
3. Calculer l'importance groupée pour chaque variable fonctionnelle ;
4. Éliminer la variable fonctionnelle la moins importante ;
5. Répéter 2 à 4 tant qu'il reste des variables disponibles.

Cet algorithme élimine récursivement les variables fonctionnelles les moins importantes (au sens du critère d'importance groupée) et sélectionne celles qui minimisent l'erreur de prédiction évaluée à l'étape 2.

Cette procédure est appliquée au problème de l'analyse des données des enregistreurs de vol pour la prédiction des risques opérationnels en aéronautique. Nous avons accès à un grand nombre de données enregistrées chaque seconde et ce durant tout le vol. L'analyse des données de vol constitue un réel défi pour les compagnies aériennes. Cela leur permet *in fine* d'avoir des mesures objectives du niveau de risque et donc de garantir un niveau de sécurité élevé. Nous analysons ainsi ces données fonctionnelles multivariées pour prédire le risque d'atterrissage long, risque auquel les compagnies aériennes sont fréquemment confrontées.

# Bibliographie

[1] Breiman, L. (2001), *Random Forests*, Machine Learning, Vol. 45, pp 5–32.
[2] Gregorutti, B., Michel, B. and Saint Pierre, P. (2014) *Grouped variable importance with random forests and application to multivariate functional data analysis*, arXiv :1411.4170.
[3] Ramsay, J. and Silverman, B. W. (2005), *Functional data analysis*, Springer Series in Statistics, Springer.
[4] Yuan, M. and Lin, Y. (2006). *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society, Series B 68 :49–67.
[5] Zhu, R., Zeng, D. and Kosorok, M. R. (2012), *Reinforcement learning trees*, The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series.