

Regression tree for functional response : application in oceanology

David Nerini^(a)

Badih Ghattas^(b)

*(a) Centre d'Océanologie de Marseille UMR LMGEM 6117 CNRS Campus de Luminy, Case 901 13288 MARSEILLE Cedex 09 e-mail: nerini@com.univ-mrs.fr
tel : (+33) 491 829 125*

(b) Institut de Mathématiques de Luminy, Campus de Luminy, Case 907 13288 MARSEILLE Cedex 09 e-mail: ghattas@lumimath.univ-mrs.fr tel : (+33) 491 829 089

Abstract

We consider here the problem of building a regression tree when the response variable is a curve. Following the work of Yu and Lambert (1999), we give a detailed analysis of an extension of regression trees to the functional case. The conditions required for the criterion to be used to construct the tree model in the functional context is discussed. This extension is applied to an oceanological problem where the objective is to predict the shape of salinity profiles using several explanatory environmental variables. Functional PCA is proposed in order to enlighten the interpretation of the tree-based model. Finally, bagging procedure allows to increase the accuracy of the functional regression tree and leads to a more stable model although the tree structure is lost.

Key words: functional regression tree, bagging, functional PCA, Legendre polynomials, aggregated model

1 Introduction

Many substantial recent works have been carried out in functional data analysis involving various fields of applied statistics (see Ramsay and Silverman, 1997, 2002 and references therein for a general overview). Particular attention has been focused on the regression of a real response variable Y given some functional predictive variable X , both in a parametric and a non-parametric framework (Cardot *et al.*, 1999, Ferraty and Vieu, 2004). The reader can have

a good overview in Ferraty (2003) about the different approaches and recent advances in functional data analysis.

In this paper, we seek to predict a functional response variable Y given a p -dimensional predictive random vector \mathbf{X} whose components are real or discrete or both. In other terms, our aim is to estimate a regression procedure $f(\cdot) = E(Y/\mathbf{X} = \cdot)$ where Y is a random function taking values in some space of infinite dimension. The estimation of $f(\cdot)$ is tackled through an extension of regression trees (CART, Breiman *et al.*, 1984, Hastie *et al.*, 2001).

Regression trees are statistical models concerned with the prediction of a real response variable Y given a set of real predictive variables \mathbf{X} . Starting from a set of n *i.i.d.* observations $\{y_i, \mathbf{x}_i\}_{i=1..n}$ of $(Y, \mathbf{X}) \in \mathbb{R} \times \mathbb{R}^p$, they are constructed by partitioning the \mathbf{X} space into a set of hypercubes and fitting a simple model (a constant) for Y in each of these regions. The model is in the form of a binary tree containing q terminal nodes (the regions r) where a predicted value of Y is given. The construction of a tree-based model, the way to select the splits and the measure of the accuracy are achieved using the following criterion called *deviance* defined for each node r of the tree as:

$$\widehat{R}(r) = \sum_{\mathbf{x}_i \in r} (y_i - \bar{y}_r)^2 \quad (1)$$

where \bar{y}_r is the average of the observations of Y belonging to the node r . These models have been widely studied in machine learning and applied statistics and present many advantages like their representation in a form of a binary tree, working in high dimension and variable ranking.

The extension of regression trees to a multivariate response variable has been first proposed by Segal (1992) for the case of longitudinal data. He considered the observations of the response variable Y , a curve sampled on an equally spaced grid, as a vector $\mathbf{y}_i = (y_{i1}, \dots, y_{im})'$. The immediate generalization of the criterion used to select the splits came with :

$$\widehat{R}(r) = \sum_{\mathbf{x}_i \in r} (\mathbf{y}_i - \bar{\mathbf{y}}_r)' \Sigma_r^{-1} (\mathbf{y}_i - \bar{\mathbf{y}}_r)$$

where $\bar{\mathbf{y}}_r$ is the $m \times 1$ vector of response means for observations within the node r and Σ_r the covariance matrix of the responses for region r . The main difficulty arising with this extension of the deviance is that we must ensure that when splitting a node r in two subnodes r_1 and r_2 , the total within subnodes deviance must be lower than the parent node deviance, that is :

$$\widehat{R}(r) \geq \widehat{R}(r_1) + \widehat{R}(r_2)$$

To satisfy this constraint, the author assumed that the covariance of the observations within the parent node, is the same as the covariance of the observations within the both subnodes. Besides this assumption, he assumed also that each realization of Y may be modeled by an autoregressive AR(1) model. This second assumption, although hard, simplifies considerably the form of the covariance matrix, and so its estimation within each node.

The second approach provided by Yu and Lambert (1999) consists in applying a principal component analysis on the response matrix before constructing the tree. Retaining a suitable number of principal components, a tree model is fitted, taking advantage of the dimensionality reduction and of the filtering step provided by the PCA. By this way, the stability of the predictor is improved. Here, the criterion used to build the model is given by :

$$\hat{R}(r) = \sum_{\mathbf{x}_i \in r} (\zeta_i - \bar{\zeta}_r)' (\zeta_i - \bar{\zeta}_r)$$

where ζ_i denotes a principal components response vector for which the dimension is attached to the number of principal components retained for the analysis and $\bar{\zeta}_r$ is the mean response vector of observations belonging to r .

With sampled curves, both these approaches, may suffer from two major drawbacks. Linear time series analysis involves strong hypothesis such as stationarity which are difficult to check especially when dealing with response variables strongly depending on the available predictive variables. The PCA approach is highly related to the sampling strategy because the estimation of the covariance matrix of the responses depends on the location of the sampling points. Moreover, both methods do not account for situations where the response variable arises as a set of sampled curves on an unequally sampling grid.

In the functional data context, this raises the difficulty on how to measure the distances between observed curves in order to evaluate the splitting criterion. For this purpose, Yu and Lambert (1999) proposed another interesting way to construct trees. They presented each response curve as a linear combination of smoothing spline basis functions and grew the tree using the coefficients of this expansion. This leads to estimate the criterion as :

$$\hat{R}(r) = \sum_{\mathbf{x}_i \in r} (\mathbf{c}_i - \bar{\mathbf{c}}_r)' \Phi (\mathbf{c}_i - \bar{\mathbf{c}}_r)$$

where \mathbf{c}_i is the vector of coefficients of the expansion into the spline basis, $\bar{\mathbf{c}}_r$ the mean vector of coefficients belonging to region r , and Φ the matrix of the inner products of the spline basis vectors. This method settles the problem of irregular sampling but may be computationally expensive.

Following this idea, we present a unified approach available to the general context of constructing regression trees when the response variable is a curve. This study focuses on the properties required by the splitting criterion to build a tree-based model.

The outline of our paper is the following. The next section reminds the main steps of a tree building procedure and is extended to the prediction of a functional response variable. We give some remarks on the estimation and the properties required by the criterion to achieve the construction of a tree. In section 2 we propose to illustrate the predictive capabilities of the functional tree on simulated data sets. Next, a case study is conducted on real data set where functional PCA (Ramsay and Silverman, 1997) is used to improve the interpretation of the results. Finally, we build an aggregated model in order to test the stability of the tree. The last section gives the conclusions.

2 Functional regression trees

Let $L = \{(y_i, \mathbf{x}_i), i = 1, \dots, n\}$ be a learning sample of n *i.i.d.* observations such that each y_i is an observation of the functional random response variable Y valued in an Hilbert space on support $[a, b] \subset \mathbb{R}$, $\mathcal{H} = \mathbb{L}^2([a, b])$. Let $\langle f, g \rangle$ denote the usual inner product of functions f and g in \mathcal{H} , defined by :

$$\langle f, g \rangle = \int_a^b f(t) g(t) dt$$

and let $\|\cdot\|$ denote the norm attached to this inner product. The p -variate vectors $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})$ are observations of the random vector $\mathbf{X} = (X_1, \dots, X_p)$ that we suppose, for the sake of simplicity, belonging to \mathbb{R}^p .

Our aim is to estimate the value of the response function Y , for an observation of \mathbf{X} , by designing a regression function on L :

$$f_L(\mathbf{x}) = E(Y/\mathbf{X} = \mathbf{x})$$

where this regression function has the form :

$$f_L(\mathbf{X}) = \sum_{j=1}^q f_j I(\mathbf{X} \in r_j)$$

where f_j is a function belonging to \mathcal{H} , the regions r_j , $j = 1, \dots, q$ are \mathbb{R}^p -polytopes with boundaries parallel to the axis and form a partition of \mathbb{R}^p , q is

the number of regions and $I(\cdot)$ is the indicator function.

The construction of the model is done sequentially. Starting from the whole set L , the space \mathbb{R}^p is split in two regions where an estimation of the function Y is given. The split is realized using a binary rule over one component of \mathbf{X} and optimizing a criterion constructed with the y_i 's. The splitting procedure is repeated recursively on each of the two regions until some stopping rules are applied. At the end, the space of the predictive variables is partitioned into q regions r_1, r_2, \dots, r_q where q is not fixed in advance. A set of binary decision rules (the boundaries of each r_j) and a predictive curve f_j are associated to each of the terminal nodes r_j . This model can be displayed as a binary tree T , containing a set $\tilde{T} = \{r_1, r_2, \dots, r_q\}$ of terminal nodes (see Fig. 3 as an example).

The predictor $f_L(\mathbf{X})$ must accurately predict any new test sample. The accuracy of a functional tree-based model is measured as usual by the true mean integrated squared error :

$$R(T) = E \|Y - f_L(\mathbf{X})\|^2 \quad (2)$$

This measure is directly related to the criterion optimized for the choice of the splitting rule at each node of the tree. The heterogeneity of each node is measured by the deviance of the observations it contains :

$$R(r) = \sum_{\mathbf{x}_i \in r} \|y_i - f_L(\mathbf{x}_i)\|^2 \quad (3)$$

2.1 The splitting criterion

In this section we give details on the estimation of the deviance used to compute the splitting criterion.

2.1.1 Node deviance and tree performance

The deviance within a node r having n_r observations (3) is estimated as in the univariate case using the average function $\bar{y}_r = \frac{1}{n_r} \sum_{\mathbf{x}_i \in r} y_i$ of the responses within the node :

$$\hat{R}(r) = \sum_{\mathbf{x}_i \in r} \|y_i - \bar{y}_r\|^2 \quad (4)$$

For a given tree T , the total deviance (2) is estimated by :

$$\widehat{R}(T) = \frac{1}{n} \sum_{r \in \widetilde{T}} \widehat{R}(r) = \frac{1}{n} \sum_{r \in \widetilde{T}} \sum_{\mathbf{x}_i \in r} \|y_i - \bar{y}_r\|^2 \quad (5)$$

Then, the estimator $\widehat{f}_L(\mathbf{X})$ of $f_L(\mathbf{X})$ that minimizes $\widehat{R}(T)$ is written :

$$\widehat{f}_L(\mathbf{X}) = \sum_{r \in \widetilde{T}} \bar{y}_r I(\mathbf{X} \in r)$$

As the tree T is grown using the observations in L , the predictor $\widehat{f}_L(\mathbf{x})$ assigns a prediction function \bar{y}_r to new observation \mathbf{x} falling through the tree and reaching the region r . The estimator given by (5) is known to be over-all optimistic because it's estimated over the learning data set L . It can be replaced by a test sample estimate :

$$\widehat{R}_{TS}(T) = \frac{1}{v} \sum_{r \in \widetilde{T}} \sum_{\mathbf{x}_j \in L_{TS}} \|y_j - \widehat{f}_L(\mathbf{x}_j)\|^2 \quad (6)$$

where $L_{TS} = \{(y_j, \mathbf{x}_j), j = 1, \dots, v\}$ is the test sample following the same distribution as the observations in the learning sample L .

If we do not have a test sample at hand, it's possible to resort to a cross-validation estimate of $R(T)$.

2.1.2 Choosing the criterion

Starting with the whole sample L , consider a splitting variable X_j and a threshold s on this variable. Define the regions $r_1 = \{i = 1, \dots, n | x_{ij} \leq s\}$ and $r_2 = \{i = 1, \dots, n | x_{ij} > s\}$ such that $r = r_1 \cup r_2$ and $r_1 \cap r_2 = \emptyset$. The within node sum of squares can be calculated in each of these half hyperplanes :

$$\widehat{R}(r_1) = \sum_{\mathbf{x}_i \in r_1} \|y_i - \bar{y}_{r_1}\|^2 \text{ and } \widehat{R}(r_2) = \sum_{\mathbf{x}_i \in r_2} \|y_i - \bar{y}_{r_2}\|^2$$

For any split s belonging to the set S of all the candidate splits, r is subdivided into r_1 and r_2 . Let $\Delta \widehat{R}(s, r) = \widehat{R}(r) - [\widehat{R}(r_1) + \widehat{R}(r_2)]$.

The selected split s^* of r into r_1 and r_2 is the split which most decreases $\widehat{R}(r)$:

$$\Delta \widehat{R}(s^*, r) = \max_{s \in S} \Delta \widehat{R}(s, r)$$

Thus, a functional regression tree is constructed by iteratively splitting the nodes in order to maximize the decrease in the heterogeneity $\widehat{R}(r)$, as in the univariate case. The decrease in $\widehat{R}(r)$ when splitting a region r into r_1 and r_2 is guaranteed because the following property is checked for any r :

$$\widehat{R}(r) = \widehat{R}(r_1) + \widehat{R}(r_2) + \frac{n_1 n_2}{n_1 + n_2} \|\bar{y}_{r_1} - \bar{y}_{r_2}\|^2 \quad (7)$$

This property arising from the decomposition of the inertia and the Huyghens theorem, is verified because the criterion $R(r)$ is a sum of squared distances. However, it can be advisable to introduce other criteria to measure the relations between individuals belonging to a region r . For instance, if the response variable Y is a density on support $[a, b]$, the measure of heterogeneity into a region r should be the generalized Kullback-Liebler divergence :

$$R(r) = \sum_{\mathbf{x}_i \in r} \sum_{\mathbf{x}_j \in r} K(y_i, y_j)$$

with :

$$K(y_i, y_j) = \int_a^b (y_i(t) - y_j(t)) \ln \frac{y_i(t)}{y_j(t)} dt$$

For any node r split into r_1 and r_2 , it still verifies :

$$R(r) \geq R(r_1) + R(r_2)$$

because of the concavity of the divergence. In this case, the prediction error of the tree is estimated as in (5) and (6).

2.2 Approximation of the \mathbb{L}^2 -distances

The problem is now reduced to the calculus of (5). In most cases, data arise from the observation at discrete points of a set of n curves. The classical approach developed further is to choose a finite-dimensional basis and to find the best projection of each sampled curve onto this basis. We consider here the case where y_i can be decomposed in terms of linear combinations of known basis functions ϕ_1, \dots, ϕ_m :

$$y_i(t) = g_i(t) + \varepsilon_i(t)$$

$$\text{where } g_i(t) = \sum_{j=0}^m c_{ij} \phi_j(t)$$

The function $g_i(t)$ represents the predicted part of the sampled function $y_i(t)$, m is the number of basis functions and $\varepsilon_i(t)$ is a residual variation that can be regarded as noise.

We now consider the sample L arising as a set of n couples (g_i, \mathbf{x}_i) , that is :

$$L = \{(g_i, \mathbf{x}_i), i = 1, \dots, n\}$$

As g_i is a linear combination of basis functions, the predicted function \bar{y}_r may be approximated by :

$$\bar{y}_r \simeq \bar{g}_r = \sum_{j=1}^m \bar{c}_j \phi_j(t)$$

where $\bar{c}_1, \dots, \bar{c}_m$ are the mean values of the column coefficients of the n_r functions belonging to node r . Using this approximation in (4) leads to :

$$\hat{R}(r) \simeq \sum_{\mathbf{x}_i \in r} \|g_i - \bar{g}_r\|^2 = \text{tr}(C_r' C_r \Phi)$$

where :

- C_r is the centred $n_r \times m$ coefficients matrix of the n_r functions belonging to node r , C_r' denotes its transpose,

- Φ is a $m \times m$ symmetric matrix which defines the metric associated to the choice of the functional basis ϕ_1, \dots, ϕ_m . It has entries $\Phi_{ij} = \langle \phi_i, \phi_j \rangle$. For instance, using orthonormal basis gives $\Phi = I_m$ where I_m denotes the m -identity matrix. In other cases, it is possible to resort to numerical quadrature or integration to evaluate Φ (Ramsay and Silverman, 1997).

Finally, for the entire tree :

$$\hat{R}(T) = \frac{1}{n} \sum_{r \in \tilde{T}} \hat{R}(r) \simeq \frac{1}{n} \sum_{r \in \tilde{T}} \text{tr}(C_r' C_r \Phi)$$

In the equation (7) the distance $\|\bar{y}_{r_1} - \bar{y}_{r_2}\|^2$ is straightforwardly approximated by $\|\bar{\mathbf{c}}_{r_1} - \bar{\mathbf{c}}_{r_2}\|_{\Phi}^2$ where $\|\cdot\|_{\Phi}$ is the norm in \mathbb{R}^m for the Φ -metric, the vector

$\bar{\mathbf{c}}_r = (\bar{c}_1, \dots, \bar{c}_m)'_r$ contains the mean coefficients of the functions belonging to r . The initial functional problem is reduced to a tractable multivariate problem. This shows how the construction of a tree using functional data is equivalent to multivariate regression tree with a particular choice of metric matrix Φ . This Φ -metric allows to consider the dependencies of different values of a sampled curve, so called *horizontal dependencies*.

3 Simulations and application.

This section is devoted to the computational aspects of a functional tree construction. Using simulations, we show that the extension to functional regression trees described in the previous section may retrieve the right model even in the presence of significant noise in the data. An application to real data set is studied.

3.1 Simulated data

Consider the following data set :

- $var1, var2, var3$ and $var4$ are four real vectors of size n , with entries randomly sampled from an uniform distribution on $[-1; 1]$. They form the set of predictive variables.
- The observations $y_i, i = 1, \dots, n$ of the functional response variable Y raises as a set of n sampled noisy curves ranging from $-\pi$ to π on a grid $\{t_1, \dots, t_K\}$ of unequally spaced points. They form the response matrix of size $n \times K$.

The n sampled functions $y_i(t_k), k = 1, \dots, K$ depend only depend on $var3$ and $var4$, following the model :

$$\left\{ \begin{array}{ll} \text{I} & y(t_k) = \sin(t_k) + \varepsilon(t_k) \quad \text{if } var3 < 0.5 \text{ and } var4 < 0.5 \\ \text{II} & y(t_k) = \cos(t_k) + \varepsilon(t_k) \quad \text{if } var3 < 0.5 \text{ and } var4 > 0.5 \\ \text{III} & y(t_k) = \frac{1}{2} \sin\left(\frac{1}{2}t_k\right) + \varepsilon(t_k) \quad \text{if } var3 > 0.5 \text{ and } var4 < 0.5 \\ \text{IV} & y(t_k) = \frac{1}{2} \cos\left(\frac{1}{2}t_k\right) + \varepsilon(t_k) \quad \text{if } var3 > 0.5 \text{ and } var4 > 0.5 \end{array} \right. \quad (8)$$

where ε is a centered Gaussian noise with variance σ^2 . This model can be displayed in the form of a binary decision tree (Fig. 1). We have fixed the values $n = 200, K = 128$ and $\sigma^2 = 1$.

We choose to project each sampled curve onto a Fourier basis such that:

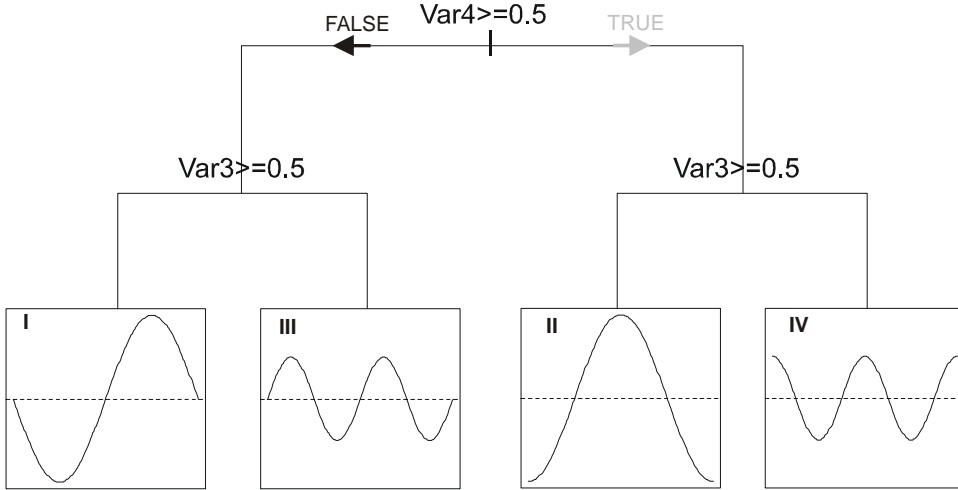


Fig. 1. Theoretical functional decision tree. Four functions are conditioned by a sequence of nested rules constructed with 4 explicative variables. The sequence of binary rules only concerns var3 and var4.

$$y_i(t) = a_0 + \sum_{j=1}^m a_j \sin \omega_j t + \sum_{i=1}^m b_j \cos(\omega_j t) + \varepsilon_i(t)$$

where the parameter ω determines the period $2\pi/\omega$ which is equal here to 2π and ε_i is a residual variation. The estimation of the $2m + 1$ coefficients $a_0, a_1, b_1, \dots, a_m, b_m$ is achieved by least square fitting (Ramsay and Silverman, 1997, p. 44). The Φ -metric is expressed as $\phi' \phi$ where ϕ is a $n \times (2m + 1)$ matrix of basis function values at the observation points with entries $\phi_j(t_k)$. We fixed the number of coefficients to five ($m = 2$) but higher values can be considered with the condition $2m + 1 \leq n$. Results of the fit are displayed on figure 2.

A learning sample L of size $n = 200$ is constituted with a 200×5 matrix of response coefficients and with the four predictive variables. A tree T is constructed and pruned with 10-fold cross-validation.

The optimal tree (3) has four terminal nodes and matches perfectly with the true model (8). The splitting rules are retrieved and the model classifies correctly the entire set of observations. The same experience has been conducted by constructing the tree on the $n \times K$ response matrix with the usual Euclidean distance. Results showed that even if the splitting rules are well found, the structure of the tree is more complex than the functional one, having much more terminal nodes. Moreover, the learning was computationally more expensive.

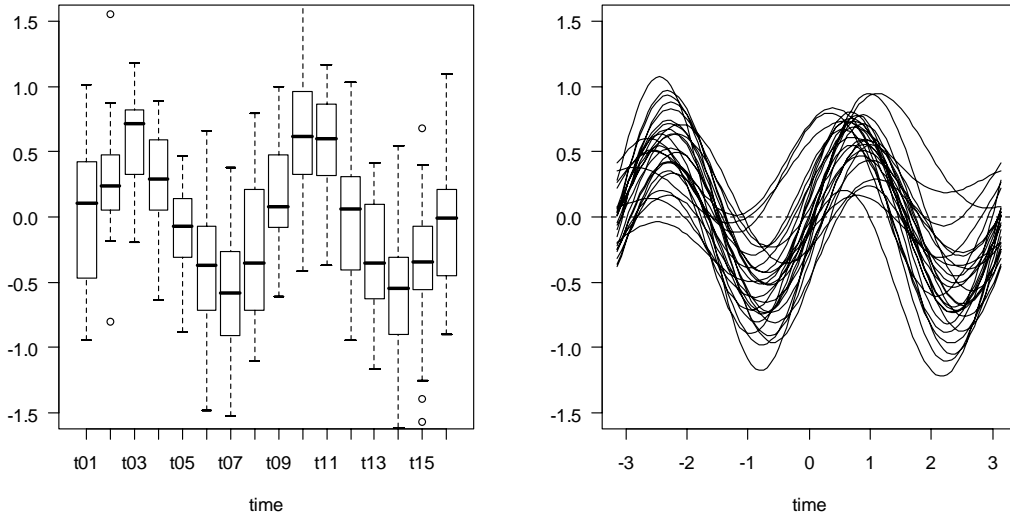


Fig. 2. Random generation of functions in node *III* with $\sigma^2 = 1$. This graphic shows on the left panel, the discrete sampling on $k = 16$ points and the expansion of each sampled curve into a Fourier basis of dimension 5.

3.2 Case study in oceanology

We seek to predict the shape of salinity curves in a shallow marine lagoon, submitted to strong perturbations such as freshwater inputs ejected from an hydroelectric power station. These freshwater inputs contribute to the installation of a vertical salinity stratification of the water column that prevents oxygen exchanges between the anoxic bottom seawater layer and the oxygenated brackish surface layer. This anoxia involves a global fall of the biodiversity of the ecosystem (Nerini *et al.*, 2000).

We want to understand the relations between the salinity profiles and the processes implied into the oxygen transfer from the bottom to the surface. In other words, we hope to predict the shape of the salinity profiles using bottom oxygen measures, freshwater flows and wind speed as predictors.

We dispose of $n = 7332$ hourly observations of salinity profiles measured at sparsely vertical discrete points on an unequally spaced time grid from 1997 to 1999. The salinity has no dimension. As a reference, the mean salinity in the Mediterranean see is 38. Each sampled curve has been estimated using a Legendre orthonormal polynomials approximation (Gautschi, 2003) (Fig. 4).

An expansion into a Legendre polynomial basis of 8 coefficients is sufficient to extract the main features of the data.

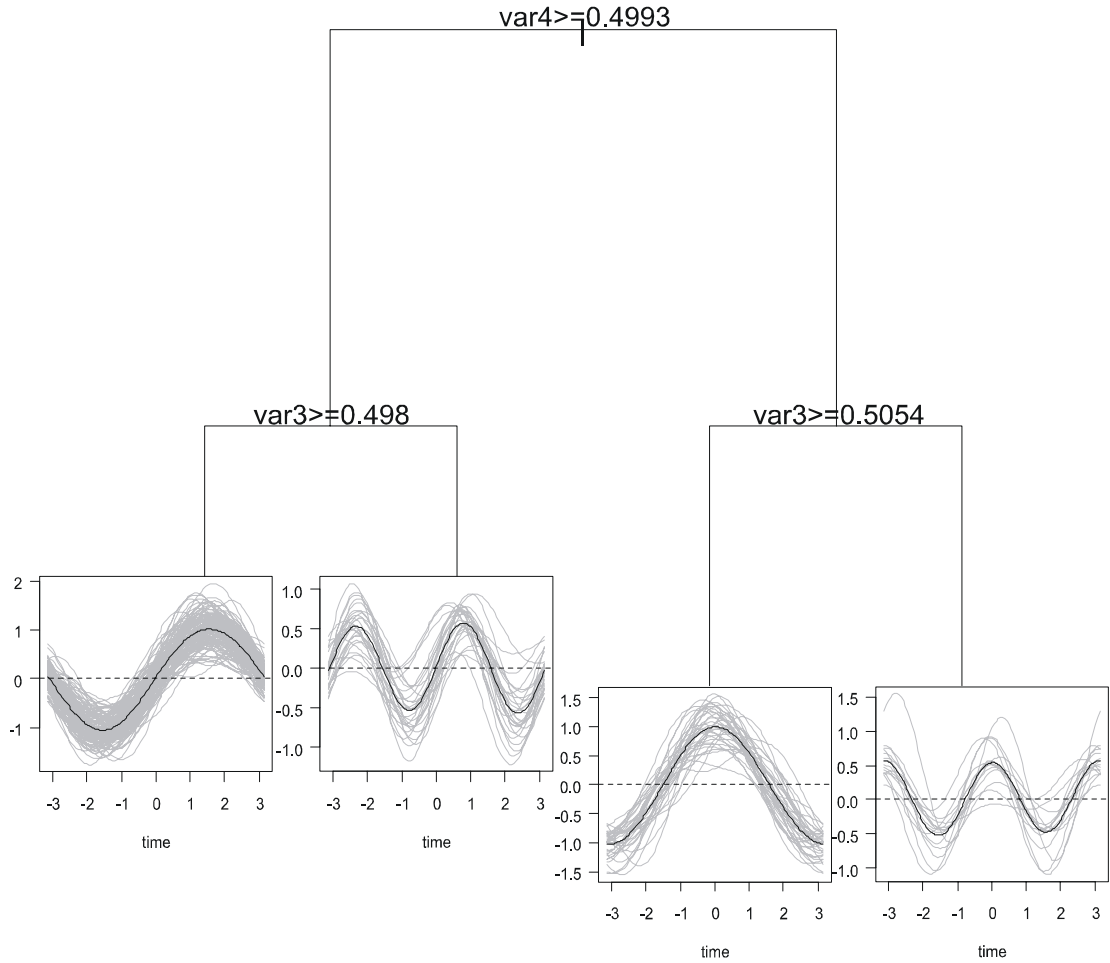


Fig. 3. Functional regression tree with simulated data ($n = 200$, $k = 128$, $m = 2$, $\sigma^2 = 1$). On this example, each observations of Y (grey curves) is well classified. The black curves represent the predicted functions in each terminal nodes. As expected, var1 and var2 don't appear into the splitting rules. The overall prediction error gives $\hat{R}(T) = 1.21$.

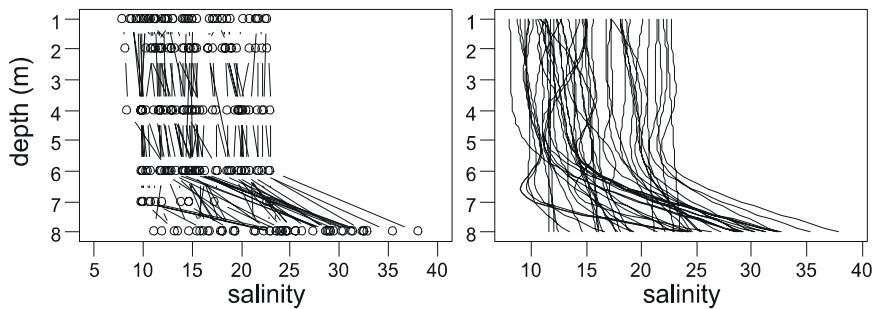


Fig. 4. Sample of 50 salinity profiles and corresponding Legendre polynomial expansion. The observations of Y are displayed as vertical curves indexed by the depth (m).

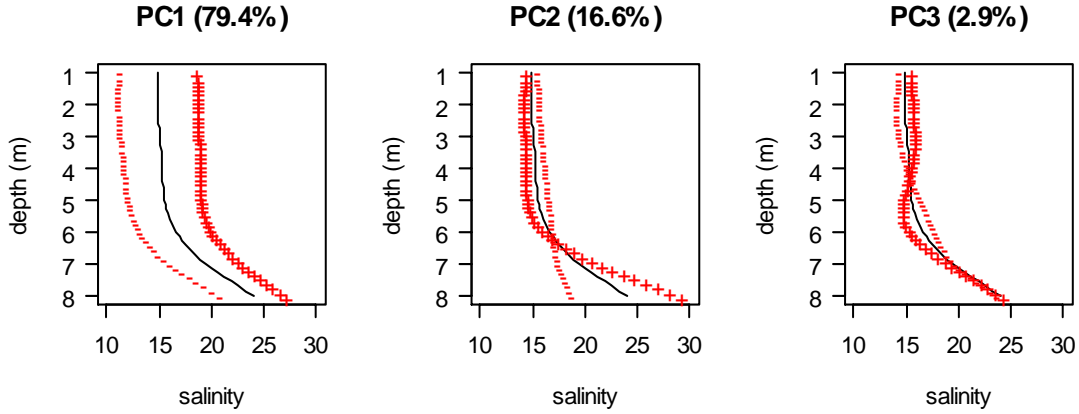


Fig. 5. Display of the first three principal components of the functional PCA of the dataset. The curves show the mean function (black curve) and the effects of adding (+) and substrating (-) a multiple of each eigenfunction. Results show that the structure of water column salinity is highly structured by the average. Few variability is accounted for the shape variations of profiles (PC2,PC3).

The set L is then constituted by a response matrix of $n \times 8$ coefficients of the polynomial expansion and a predictive $n \times 3$ matrix of the explanatory variables : the maximum wind speed (m/s), the freshwater flow integrated over the last 15 days ($m^3/s/15days$) and the dissolved oxygen rate measured at the bottom (in percentages).

As it can be seen on figure 5, three different sources of variability accounting for 99% of the total variance, are identified with functional PCA computed on the polynomial coefficients of the expansion. If we denote as \bar{g} the overall mean function of the curves $g_i, i = 1, \dots, 7332$, the effect of the eigenfunction ξ_j associated to eigenvalue λ_j of the PCA can be displayed as $\bar{g} \pm \sqrt{\lambda_j} \xi_j$ (Ramsay and Silverman, 1997).

The first factor of variability involves a translation of the profile over the range of salinity. The salinity is in fact strongly attached to the season where sampling took effect : we call it *mean effect*. High mean values of salinity curves are encountered in summer, lower values characterize winter samples when the amount of freshwater released is high. This great part of the variability demonstrates that the presence of the stratification doesn't depend on the season. The second source identifies shape variations of salinity. Profiles are either straight from the surface to the bottom or they present a high curvature near six meters, which confirms the presence of a strong stratification. The last factor shows opposite movements of the salinity profile from the surface to six meters.

A 10-fold cross-validated tree is grown using the whole set L restricting the terminal nodes to have at least 200 observations (Fig. 6). In order to enlighten

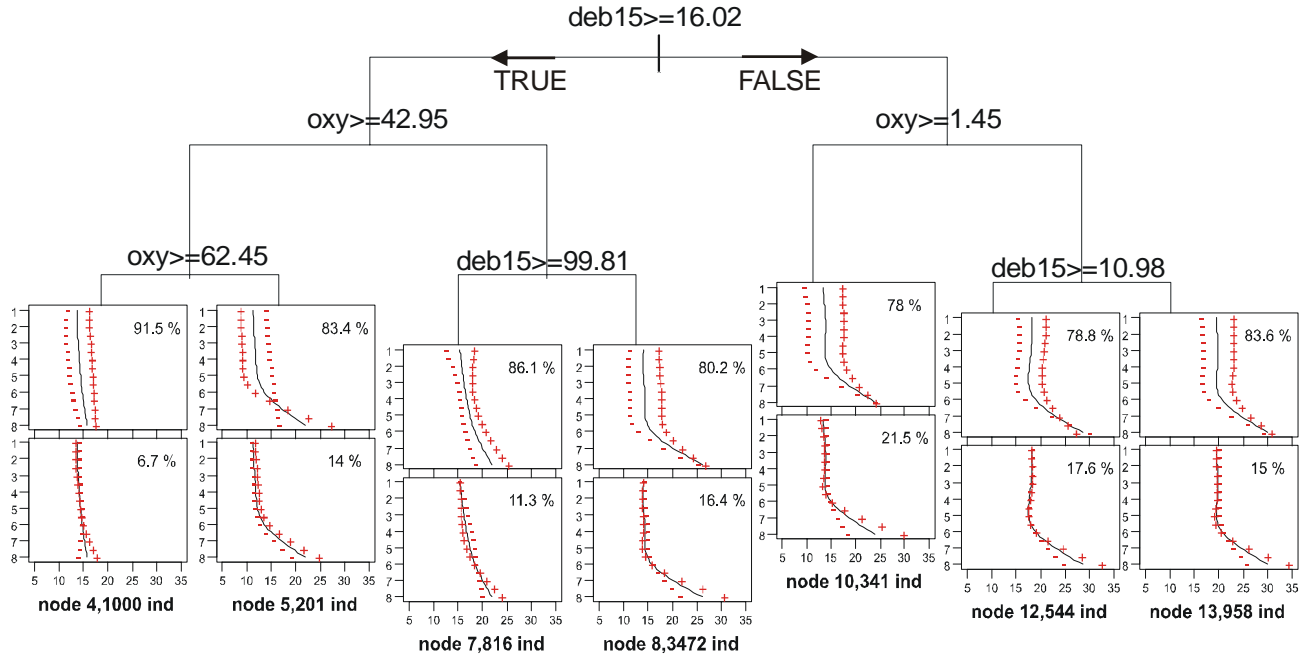


Fig. 6. Functional regression tree constructed over 7332 observations of salinity profiles. Only 2 predictive variables appeared on this 3-splits-level tree. A terminal nodes contains the number of observations and the predictive curve (black line). In each terminal node, the result of the local PCA with the % of variability is displayed for the 2 first components accounting for often more than 90% of the variability.

the relations between observations into a terminal node we performed a local functional PCA. The effects of the first and the second eigenfunctions are both displayed in each terminal node, containing the number of individuals and the percentage of variability accounting for each score. For instance, the extreme left terminal node 4 exhibits a vertical predicted profile \bar{y}_4 (black curve) associated to high values of oxygen. The main part of the variability (91.5%) in this node is essentially due to the mean effect (translated curves).

Interpreting the tree, the freshwater flow is the most important variable that structure the system. Its influence involves essentially a mean effect of the profiles, starting from vertical profiles on the left to high stratified one on the right. The left and right parts of the tree oppose winter observations to summer (high freshwater values to low freshwater values) The stratification is always present in summer (right), when high values of freshwater are strongly linked to the salinity profiles in winter. Node 5 shows an interesting feature of salinity profiles enlightened by the PCA. The predicted function is a stratified profile and yet vertical profiles and stratified profile are mixed in the same node. This shows that the stratification cannot be completely associated to low values of oxygen. The entire tree can be interpreted in the same way.

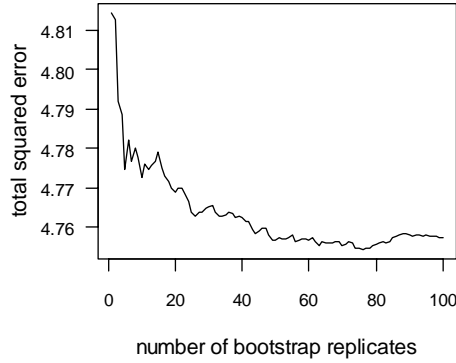


Fig. 7. Improvement of predictive performances of the functional regression model using bagged predictors. As the number of aggregated predictor increases the total squared error decreases.

3.3 Aggregating functional trees.

The main drawback of regression trees is that slight modifications of the learning sample can lead to drastically different structure of the model. To test the robustness of the tree structure, we construct an aggregated predictor by combining multiple versions of the model constructed on bootstrap samples of L so called *bagging* (Breiman, 1996). The bootstrap samples $L_b, b = 1, \dots, B$ are replicated data sets each consisting of n individuals randomly drawn from L with replacement. Over each bootstrap sample, we construct a functional predictor $\hat{f}_{L_b}(\cdot)$ and give the bagged prediction $\hat{f}_{bag}(\mathbf{x})$ for a new observation \mathbf{x} as :

$$\hat{f}_{bag}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^B \hat{f}_{L_b}(\mathbf{x})$$

Before constructing the aggregated predictor, the data set is randomly partitioned in two data sets : L_{TS} , a test set containing 10% of the data and L , the learning set with 90% of the remaining data. The aggregated predictor accuracy is evaluates on the test sample L_{TS} .

Using 100 bootstrap samples (Fig. 7) the prediction error is computed from the bagged predictor over the test sample. For this case study, we gain in prediction accuracy. The improvement may appear to be slight, probably because the trees constructed over the bootstrap samples are optimized by cross-validation.

4 Conclusion.

In the context of predicting a functional response variable with regression trees, we have chosen to represent the sampled response curves in a finite-dimensional basis and to find the best projection of each curve onto this basis. This transformation leads to a nice approximation of the generalized splitting criterion reducing the functional problem to a simpler multivariate problem. Finally, we have illustrated with numerical simulations that the functional regression tree approach captures the true model even in the presence of significant noise on the responses. For the case study considered here, we got a satisfactory predictive model where the shape of the responses is well-classified using the available explanatory variables. We improved the prediction accuracy with bagging.

Our approach relies on the choice and the dimension of the projection basis. The choice of the basis may be guided by the shape and the properties of the response curves data set. However, the number of basis functions is strongly related to the number and the location of sampling points available for each curve especially when coefficients are estimated by least squares.

A convenient alternative is first to apply a non-parametric smoothing to estimate the unknown response curves. In case of polynomials, the calculus of $\|y_i - \bar{y}_r\|^2$ may be achieved thanks to quadrature rules. The problem of sparsely sampled curves is thus tackled in the following way. First, each curve is estimated with non-parametric regression on the discrete observations. Then, we take quadrature points for which their location depends on the selected quadrature scheme, using the nonparametric estimated curve as support. Finally, a tree is directly build on these points used as response variable. With this method, and for sufficiently regular curves, the computational time is improved regarding the number of basis coefficients. Moreover, this number can be higher than the number of the initial sampled points and the problem of curves sampled on unequally spaced grid is fixed as well.

References

- [1] Breiman, L., Friedman, J.H., Olshen, R. and Stone C.J. (1984). *Classification And Regression Trees*. Wadsworth, Belmont CA.
- [2] Breiman, L. (1996). Bagging Predictors, *Machine Learning*, **24**, 123-140
- [3] Cardot, H., Ferraty, F. and Sarda P. (1999). Functional Linear Model. *Statistics and Probability Letters*. **1**, 11-22.

- [4] Ferraty, F. (2003). *Modélisation statistique pour variables aléatoires fonctionnelles : théorie et applications*. Habilitation à diriger des recherches, Univ. P. Sabatier, Toulouse, France.
- [5] Ferraty, F. and Vieu, P. (2004). Nonparametric models for functional data, with application in regression, time series and classification. *Nonparametric Statistics*, **16**, 111-127.
- [6] Gautschi W. (2004). *Orthogonal polynomials. Computation and approximation*. Oxford Science Publications.
- [7] Hastie. T., Tibshirani R. and Friedman J. (2001). *The Elements of Statistical Learning*. Springer-Verlag.
- [8] Nerini, D., Durbec, J.P. and Mante, C. (2000). Analysis of oxygen rate time series in a strongly polluted pond. An approach by regression tree methods. *Ecol. Mod.* , **133**, 95-105
- [9] Ramsay, J. O. and Silverman, B. W. (1997). *Functional data analysis*. Springer-Verlag.
- [10] Ramsay, J. O. and Silverman, B. W. (2003). *Applied Functional Data Analysis*. Springer-Verlag.
- [11] Segal M.R. (1992). Tree-Structured Methods for Longitudinal Data. *J. Am. Statist. Ass.*, **87**, 407-418.
- [12] Yu, Y. and Lambert, D. (1999). Fitting Trees to Functional Data: With an Application to Time-of-day Patterns. *J. Comp. Graph. Stat.*, **8**, 749-762.