Random Forests for time-fixed and time-dependent predictors: The DynForest R package

Anthony Devaux, Cécile Proust-Lima, Robin Genuer

Random Forests for time-fixed and time-dependent predictors: The DynForest R package

A Preprint

Anthony Devaux *
Inserm U1219, Bordeaux Population Health
Université de Bordeaux
Bordeaux, FRANCE
anthony.devauxbarault@gmail.com

Cécile Proust-Lima
Inserm U1219, Bordeaux Population Health
Université de Bordeaux
Bordeaux, FRANCE
cecile.proust-lima@u-bordeaux.fr

Robin Genuer
Inserm U1219, Bordeaux Population Health
Université de Bordeaux
Bordeaux, FRANCE
robin.genuer@u-bordeaux.fr

February 2, 2023

Abstract

The **R** package **DynForest** implements random forests for predicting a categorical or a (multiple causes) time-to-event outcome based on time-fixed and time-dependent predictors. Through the random forests, the time-dependent predictors can be measured with error at subject-specific times, and they can be endogeneous (i.e., impacted by the outcome process). They are modeled internally using flexible linear mixed models (thanks to **lcmm** package) with time-associations pre-specified by the user. **DynForest** computes dynamic predictions that take into account all the information from time-fixed and time-dependent predictors. DynForest also provides information about the most predictive variables using variable importance and minimal depth. Variable importance can also be computed on groups of variables. To display the results, several functions are available such as summary and plot functions. This paper aims to guide the user with a step-by-step example of the different functions for fitting random forests within **DynForest**.

***K*eywords** dynamic prediction · survival data · competing risks · regression · classification · longitudinal predictors · random forests · **R**

1 Introduction

Random forests are a non-parametric powerful method for prediction purpose. Introduced by Breiman (Breiman 2001) for classification (categorical outcome) and regression (continuous outcome) frameworks, random forests are particularly designed to tackle modeling issues in high-dimension context ($n << p$). They can also easily take into account complex association between the outcome and the predictors without any pre-specification where regression models are rapidly limited.

Recently, this methodology was extended to survival data (Hemant Ishwaran et al. 2008) and competing events (Hemant Ishwaran et al. 2014). Random forests were implemented in several **R** (R Core Team 2019) packages such as **randomForestSRC** (H. Ishwaran and Kogalur 2022), **ranger** (Wright and Ziegler 2017)

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies. Optional.

or **xgboost** (Chen and Guestrin 2016) among others. However, these packages are all limited to time-fixed predictors. Yet, in many applications, it may be relevant to include predictors that are repeatedly measured at multiple occasions (regular or irregular times) with measurement errors to more accurately predict the outcome. This is the case in particular in health research where an health outcome is to be predicted according to the history of individual information.

We developed an original random forests methodology to tackle this issue and incorporate longitudinal predictors that may be prone-to-error and possibly intermittently measured (Devaux et al. 2022). The present paper aims to describe the **DynForest R** package associated to this methodology, allowing to predict a continuous, categorical or survival outcome using multivariate time-dependent predictors.

In section 2, we briefly present **DynForest** methodology through its algorithm. In section 3, we present the different functions of **DynForest** and we illustrate them in section 4 for a survival outcome, in section 5 for a categorical outcome and in section 6 for a continuous outcome. To conclude, we discuss in section 7 the limitations and future improvements.

2 DynForest principle

DynForest is a random forest methodology which can include both time-fixed predictors of any nature and time-dependent predictors possibly measured at irregular times. The purpose of **DynForest** is to predict an outcome which can be categorical, continuous or survival (with possibly competing events).

The random forest should first be built on a learning dataset of N subjects including: Y the outcome; $\mathcal{M}_x$ an ensemble of P time-fixed predictors; $\mathcal{M}_y$ an ensemble of Q time-dependent predictors. As usually, the random forest consists of an ensemble of B trees which are grown as detailed below, and aggregated to obtain the predictions.

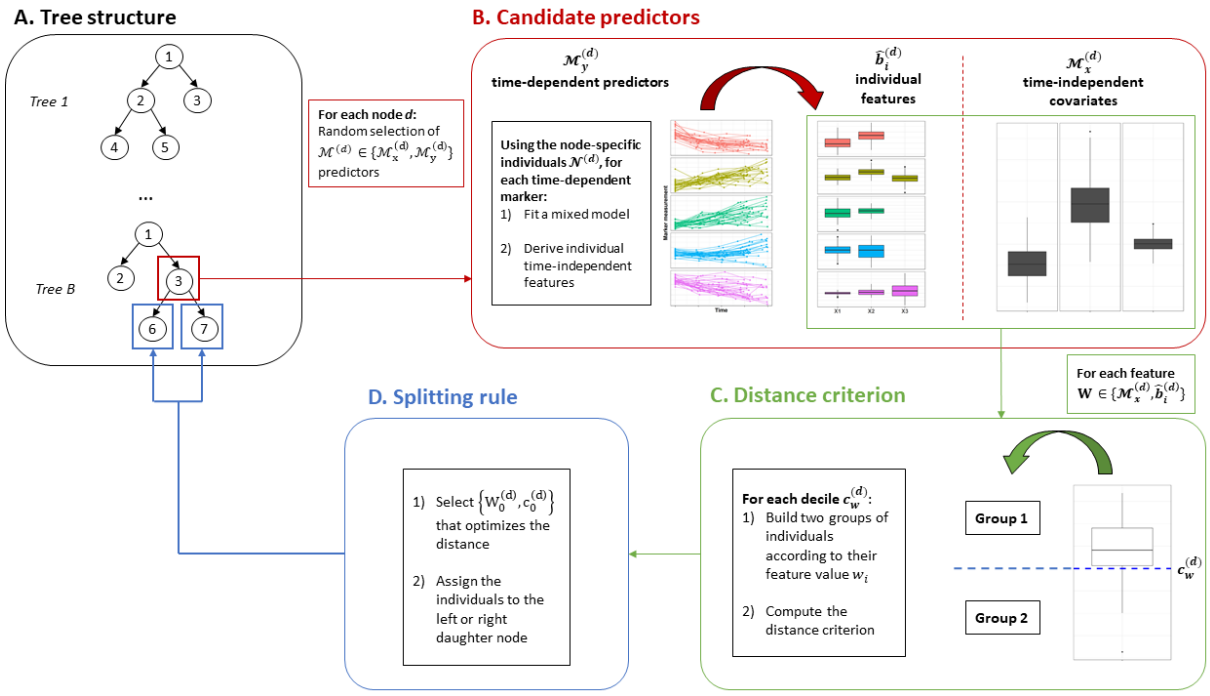

Figure 1: Overall scheme of the tree building in **DynForest** with (A) the tree structure, (B) the node-specific treatment of time-dependent predictors to obtain time-fixed features, (C) the dichotomization of the time-fixed features, (D) the splitting rule.

2.1 The tree building

The tree building process, summarized in Figure 1, aims to recursively partition the subjects into groups/nodes that are the most homogeneous regarding the outcome Y.

For each tree b ($b = 1, ..., B$), we first draw a bootstrap sample from the original dataset of N subjects (N draws among the N subjects with replacement). The subjects excluded by the bootstrap constitute the out-of-bag (OOB) sample, noted OOB^b for tree b. At each node $d \in \mathcal{D}^b$ of the tree, we recursively repeat the following steps using the $N^{(d)}$ subjects located at node d:

1. An ensemble of $\mathcal{M}^{(d)} = \{\mathcal{M}_x^{(d)}, \mathcal{M}_y^{(d)}\}$ candidate predictors are randomly selected among $\{\mathcal{M}_x, \mathcal{M}_y\}$ (see Figure 1B). The size of $\mathcal{M}^{(d)}$ is defined by the hyperparameter `mtry`.
2. For each time-dependent predictor in $\mathcal{M}_y^{(d)}$:
    a. We independently model the trajectory of the predictor using a flexible linear mixed model (Laird and Ware 1982) according to time (the specification of the model is defined by the user).
    b. We derive an ensemble $\mathcal{M}_{y\star}^{(d)}$ of individual time-independent features. These features are the individual random-effects of the linear mixed model predicted from the repeated data of individual $i = 1, ..., N^{(d)}$.
3. We define $\mathcal{M}_\star^{(d)} = \{\mathcal{M}_x^{(d)}, \mathcal{M}_{y\star}^{(d)}\}$ our new ensemble of candidate features.
4. For each candidate feature $W \in \mathcal{M}_\star^{(d)}$:
    a. We build a serie of splits $c_W^{(d)}$ according to the feature values if continuous, or subsets of categories otherwise (see Figure 1C), leading each time to two groups.
    b. We quantify the distance between the two groups according to the nature of Y:
        - If Y continuous: We compute the weighted within-group variance with the proportion of subjects in each group as weights.
        - If Y categorical: We compute the weighted within-group Shannon entropy (Shannon 1948) (i.e., the amount of uncertainty) with the proportion of subjects in each group as weights.
        - If Y survival <u>without</u> competing events: We compute the log-rank statistic test (Peto and Peto 1972).
        - If Y survival <u>with</u> competing events: We compute the Fine & Gray statistic test (R. J. Gray 1988).
5. We split the subjects into the two groups that minimize (for continuous and categorical outcome) or maximize (for survival outcome) the quantity defined previously. We denote $\{W_0^d, c_0^d\}$ the optimal couple used to split the subjects and assign them to the left and right daughter nodes $2d$ and $2d + 1$, respectively (see Figure 1D and A).
6. Step 1 to 5 are iterated on the daughter nodes until stopping criteria are met.

We define two stopping criteria: one according to `nodesize`, the minimal number of subjects in a node required to reiterate the split, and another according to `minsplit`, the minimal number of events required to split the node. `minsplit` is only defined with survival outcome. In the following, we call leaves the terminal nodes.

In each leaf $h \in \mathcal{H}^b$ of tree b, a summary π^{h^b} is computed using the individuals belonging to the leaf. The leaf summary is defined according to the outcome:

- The mean, for Y continuous.
- The category with the highest probability, for Y categorical.
- The cumulative incidence function over time computed using the Nelson-Aalen cumulative hazard function estimator (Nelson 1969; O. Aalen 1976), for Y single cause time-to-event.
- The cumulative incidence function over time computed using the non-parametric Aalen-Johansen estimator (O. O. Aalen and Johansen 1978), for Y time-to-event with multiple causes.

2.2 Individual prediction of the outcome

2.2.1 Out-of-bag individual prediction

The overall OOB prediction $\hat{\pi}_\star$ for a subject $\star$ can be computed using the tree-based predictions of $\star$ over the random forest as follows:

- with Y continuous or survival:

$$\hat{\pi}_\star = \frac{1}{|\mathcal{O}_\star|} \sum_{b \in \mathcal{O}_\star} \hat{\pi}^{h_\star^b} \tag{1}$$

- with Y categorical:

$$\hat{\pi}_\star = \underset{g \in \mathcal{G}}{\text{argmax}} \Big\{ \sum_{b \in \mathcal{O}_\star} \mathbb{1}_{(\hat{\pi}^{h^b_\star} = g)} \Big\} \tag{2}$$

where $\mathcal{G}$ indicates the ensemble of Y categories (if categorical), $\mathcal{O}_\star$ is the ensemble of trees where $\star$ is OOB and $|\mathcal{O}_\star|$ denotes its cardinal.

The prediction $\hat{\pi}^{h^b_\star}$ is obtained by dropping down subject $\star$ along tree b. At each node $d \in \mathcal{D}^b$, the subject $\star$ is assigned to the left or right node according to his/her data and the optimal couple $\{W_0^d, c_0^d\}$. W_0^d is a random-effect feature, its value for $\star$ is predicted from the individual repeated measures using the estimated parameters from the linear mixed model.

2.2.2 Individual dynamic prediction from a landmark time

With a survival outcome, the OOB prediction described in the previous paragraph can be extended to compute the individual probability of event from a landmark time s by exploiting the repeated measures of subject $\star$ only until s. For a new subject $\star$, we thus define the individual prediction $\hat{\pi}_\star(s)$ at landmark time s with:

$$\hat{\pi}_\star(s) = \frac{1}{B} \sum_{b=1}^{B} \hat{\pi}^{h^b_\star}(s) \tag{3}$$

where $\hat{\pi}^{h^b_\star}(s)$ is the tree-based prediction computed by dropping down $\star$ along the tree by considering longitudinal predictors collected until s and time-fixed predictors.

2.3 Out-of-bag error

Using the OOB individual predictions, an OOB error can be internally assessed. The OOB error quantifies the difference between the observed and the predicted values. It is defined according to the nature of Y as:

- For Y continuous, the mean square error (MSE) defined by:

$$errOOB = \frac{1}{N} \sum_{i=1}^{N} (\hat{\pi}_i - \pi_i^0)^2 \tag{4}$$

- For Y categorical, the missclassification error defined by:

$$errOOB = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{(\hat{\pi}_i \neq \pi_i^0)} \tag{5}$$

- For Y survival, the integrated Brier score (IBS) (Sène et al. 2016) between τ_1 and τ_2 defined by:

$$errOOB = \int_{\tau_1}^{\tau_2} \frac{1}{N} \sum_{i=1}^{N} \hat{\omega}_i(t) \Big\{ \mathbb{1}_{(T_i \leq t, \delta_i = k)} - \hat{\pi}_{ik}(t) \Big) \Big\}^2 dt \tag{6}$$

with π^0 the observed outcome, T the time-to-event, k the cause of interest and $\hat{\omega}(t)$ the estimated weights using inverse probability of censoring weights (IPCW) technique that accounts for censoring (Gerds and Schumacher 2006).

The OOB error of prediction is used in particular to tune the random forest by determining the hyperparameters (i.e., `mtry`, `nodesize` and `minsplit`) which give the smallest OOB error.

2.4 Explore the most predictive variables

2.4.1 Variable importance

The variable importance (VIMP) measures the loss of predictive performance (Hemant Ishwaran et al. 2008) when removing the link between a predictor and the outcome. The link is removed by permuting the predictor values at the subject level for time-fixed predictors or at observation level for time-dependent predictors, in the OOB samples. A large VIMP value indicates a good prediction ability for the predictor.

However, in case of correlated predictors, the VIMP may not properly quantify the variable importance (Gregorutti, Michel, and Saint-Pierre 2017) as the information of the predictor may still be present. To better handle situations with highly correlated predictors, the grouped variable importance (gVIMP) can be computed. It consists in simultaneously evaluate the importance of a group of predictors defined by the user. The computation is the same as for the VIMP except the permutation is performed simultaneously on all the predictors of the group. A large gVIMP value indicates a good prediction ability for the group of predictors.

2.4.2 Minimal depth

The minimal depth is another statistic to quantify the importance of a variable. It assesses the distance between the root node and the first node for which the predictor is used to split the subjects (1 for first level, 2 for second level, 3 for third level, ...). This statistic can be computed at the predictor level or at the feature level, allowing to fully understand the tree building process.

We strongly advice to compute the minimal depth with `mtry` hyperparameter chosen at its maximum to ensure that all predictors are systematically among candidate predictors for splitting the subjects.

3 DynForest R package

DynForest methodology was implemented into the **R** package **DynForest** (Devaux 2022) freely available on the comprehensive **R** archive network (CRAN) to users.

The package includes two main functions: `DynForest()` and `predict()` for the learning and the prediction steps. These functions are fully described in Section 3.1 and 3.2. Other functions available are briefly described in Table 1. These functions are illustrated in examples, one for a survival outcome, one for a categorical outcome and one for a continuous outcome.

Table 1: Brief description of the functions available in **DynForest**.

Function	Description
Learning and prediction steps	
`DynForest()`	Function that builds the random forest
`predict()`	Function that predicts the outcome on new subjects using the individual-specific information (S3 method)
Assessment function	
`compute_OOBerror()`	Function that computes the out-of-bag error to be minimized to tune the random forest
Exploring functions	
`compute_VIMP()`	Function that computes the importance of variables
`compute_gVIMP()`	Function that computes the importance of a group of variables
`var_depth()`	Function that extracts information about the tree building process
Plot function	
`plot()`	Function that plots the minimal depth by predictors or features, the importance of variables by value or percentage or the importance of a group of variables by value or percentage (S3 method)
Other function	
`summary()`	Function that displays information about the type of random forest, predictors included, parameters used, out-of-bag error (if computed) and brief summaries about the leaves (S3 method)

3.1 `DynForest()` function

`DynForest()` is the function to build the random forest. The call of this function is:

```
DynForest(timeData = NULL, fixedData = NULL, idVar = NULL,
          timeVar = NULL, timeVarModel = NULL, Y = NULL,
          ntree = 200, mtry = NULL, nodesize = 1, minsplit = 2, cause = 1,
          nsplit_option = "quantile", ncores = NULL,
          seed = round(runif(1, 0, 10000)), verbose = TRUE)
```

3.1.1 Arguments

`timeData` is an optional argument that contains the dataframe in longitudinal format (i.e., one observation per row) for the time-dependent predictors. In addition to time-dependent predictors, this dataframe should include a unique identifier and the measurement times. This argument is set to `NULL` if no time-dependent predictor is included. Argument `fixedData` contains the dataframe in wide format (i.e., one subject per row) for the time-fixed predictors. In addition to time-fixed predictors, this dataframe should also include the same identifier as used in `timeData`. This argument is set to `NULL` if no time-fixed predictor is included. Argument `idVar` provides the name of identifier variable included in `timeData` and `fixedData` dataframes. Argument `timeVar` provides the name of time variable included in `timeData` dataframe. Argument `timeVarModel` contains a list that specifies the structure of the mixed models assumed for each longitudinal predictor of `timeData` dataframe. For each longitudinal predictor, the list should contain a `fixed` and a `random` argument to define the formula of a mixed model to be estimated with **lcmm R** package (Proust-Lima, Philipps, and Liquet 2017). `fixed` defines the formula for the fixed-effects and `random` for the random-effects. Argument `Y` contains a list of two elements `type` and `Y`. Element `type` defines the nature of the outcome (`surv` for survival outcome with possibly competing causes, `numeric` for continuous outcome and `factor` for categorical outcome) and element `Y` defines the dataframe which includes the identifier (same as in `timeData` and `fixedData` dataframes) and outcome variables.

Arguments `ntree`, `mtry`, `nodesize` and `minsplit` are the hyperparameters of the random forest. Argument `ntree` controls the number of trees in the random forest (200 by default). Argument `mtry` indicates the number of variables randomly drawn at each node (square root of the total number of predictors by default). Argument `nodesize` indicates the minimal number of subjects allowed in the leaves (1 by default). Argument `minsplit` controls the minimal number of events required to split the node (2 by default).

For survival outcome, argument `cause` indicates the event the interest. Argument `nsplit_option` indicates the method to build the two groups of individuals at each node. By default, we build the groups according to deciles (`quantile` option) but they could be built according to random values (`sample` option).

Argument `ncores` indicates the number of cores used to grow the trees in parallel mode. By default, we set the number of cores of the computer minus 1. Argument `seed` specifies the random seed. It can be fixed to replicate the results. Argument `verbose` allows to display a progression bar during the execution of the function.

3.1.2 Values

`DynForest()` function returns an object of class `DynForest` containing several elements:

- `data` a list with longitudinal predictors (`Longitudinal` element), continuous predictors (`Numeric` element) and categorical predictors (`Factor` element);
- `rf` is a dataframe with one column per tree containing a list with several elements, which includes:
  - `leaves` the leaf identifier for each subject used to grow the tree;
  - `idY` the identifiers for each subject used to grow the tree;
  - `V_split` the split summary (more detailed below);
  - `Y_pred` the estimated outcome in each leaf;
  - `model_param` the estimated parameters of the mixed model for the longitudinal predictors used to split the subjects at each node;
  - `Ytype`, `hist_nodes`, `Y`, `boot` and `Ylevels` internal information used in other functions;
- `type` the nature of the outcome;
- `times` the event times (only for survival outcome);
- `cause` the cause of interest (only for survival outcome);
- `causes` the unique causes (only for survival outcome);
- `Inputs` the list of predictors names for `Longitudinal`; (longitudinal predictor), `Continuous` (continuous predictor) and `Factor` (categorical predictor)
- `Longitudinal.model` the mixed model specification for each longitudinal predictor;
- `param` a list of hyperparameters used to grow the random forest;
- `comput.time` the computation time.

The main information returned by `rf` is `V_split` element. `V_split` returns a table sorted by the node/leaf identifier (`id_node` column) with each row representing a node/leaf. Each column provides information about the splits:

- `type`: the nature of the predictor (`Longitudinal` for longitudinal predictor, `Numeric` for continuous predictor or `Factor` for categorical predictor) if the node was split, `Leaf` otherwise;
- `var_split`: the predictor used for the split defined by its order in `timeData` and `fixedData`;
- `feature`: the feature used for the split defined by its position in random statistic;
- `threshold`: the threshold used for the split (only with `Longitudinal` and `Numeric`). No information is returned for `Factor`;
- `N`: the number of subjects in the node/leaf;
- `Nevent`: the number of events of interest in the node/leaf (only with survival outcome);
- `depth`: the depth level of the node/leaf.

3.1.3 Additional information about the dependencies

`DynForest()` function internally calls other functions from related packages to build the random forest:

- `hlme()` function (from **lcmm** package (Proust-Lima, Philipps, and Liquet 2017)) to fit the mixed models for the time-dependent predictors defined in `timeData` and `timeVarModel` arguments;
- `Entropy()` function (from **base** package) to compute the Shannon entropy;
- `survdiff()` function (from **survival** package (Therneau 2022)) to compute the log-rank statistic test;
- `crr()` function (from **cmprsk** package (B. Gray 2020)) to compute the Fine & Gray statistic test.

3.2 `predict()` function

`predict()` is the function (S3 method) to predict the outcome on new subjects. Landmark time can be specified to consider only longitudinal data collected up to this time to compute the prediction. The call of this function is:

```
predict(object, timeData = NULL, fixedData = NULL,
        idVar, timeVar, t0 = NULL)
```

3.2.1 Arguments

Argument `object` contains a `DynForest` object resulting from `DynForest()` function. Argument `timeData` contains the dataframe in longitudinal format (i.e., one observation per row) for the time-dependent predictors of new subjects. In addition to time-dependent predictors, this dataframe should also include a unique identifier and the time measurements. This argument can be set to `NULL` if no time-dependent predictor is included. Argument `fixedData` contains the dataframe in wide format (i.e., one subject per row) for the time-fixed predictors of new subjects. In addition to time-fixed predictors, this dataframe should also include an unique identifier. This argument can be set to `NULL` if no time-fixed predictor is included. Argument `idVar` provides the name of the identifier variable included in `timeData` and `fixedData` dataframes. Argument `timeVar` provides the name of time-measurement variable included in `timeData` dataframe. Argument `t0` defines the landmark time; only the longitudinal data collected up to this time are to be considered. This argument should be set to `NULL` to include all longitudinal data.

3.2.2 Values

`predict()` function returns several elements:

- `t0` the landmark time defined in argument (`NULL` by default).
- `times` times used to compute the individual predictions (only with survival outcome). The times are defined according to the time-to-event subjects used to build the random forest.
- `pred_indiv` the predicted outcome for the new subject. With survival outcome, predictions are provided for each time defined in `times` element.
- `pred_leaf` a table giving for each tree (in column) the leaf in which each subject is assigned (in row).
- `pred_indiv_proba` the proportion of the trees leading to the category prediction for each subject (only with categorical outcome).

4 How to use DynForest R package with a survival outcome?

4.1 Illustrative dataset: pbc2 dataset

The pbc2 dataset (Murtaugh et al. 1994) is loaded with the package **DynForest** to illustrate its function abilities. pbc2 data come from a clinical trial conducted by the Mayo Clinic between 1974 and 1984 to treat the primary biliary cholangitis (PBC), a chronic liver disease. 312 patients were enrolled in a clinical trial to evaluate the effectiveness of D-penicillamine compared to a placebo to treat the PBC and were followed since the clinical trial ends, leading to a total of 1945 observations. During the follow-up, several clinical continuous markers were collected over time such as: the level of serum bilirubin (`serBilir`), the level of serum cholesterol (`serChol`), the level of albumin (`albumin`), the level of alkaline (`alkaline`), the level of aspartate aminotransferase (`SGOT`), platelets count (`platelets`) and the prothrombin time (`prothrombin`). 4 non-continuous time-dependent predictors were also collected: the presence of ascites (`ascites`), the presence of hepatomegaly (`hepatomegaly`), the presence of blood vessel malformations in the skin (`spiders`) and the edema levels (`edema`). These time-dependent predictors were recorded according to `time` variable. In addition to these time-dependent predictors, few time-fixed predictors were collected at enrollment: the sex (`sex`), the age (`age`) and the drug treatment (`drug`). During the follow-up, 140 patients died before transplantation, 29 patients were transplanted and 143 patients were censored alive (`event`). The time of first event (censored alive or any event) was considered as the event time (`years`)

```
library("DynForest")
head(pbc2)
```

```
##    id      time ascites hepatomegaly spiders                    edema serBilir
## 1   1 0.0000000     Yes          Yes     Yes edema despite diuretics     14.5
## 2   1 0.5256817     Yes          Yes     Yes edema despite diuretics     21.3
## 3  10 0.0000000     Yes           No     Yes edema despite diuretics     12.6
## 4 100 0.0000000      No          Yes      No                No edema      2.3
## 5 100 0.4681853      No          Yes      No                No edema      2.5
## 6 100 0.9801774     Yes           No      No      edema no diuretics      2.9
##   serChol albumin alkaline  SGOT platelets prothrombin histologic      drug
## 1     261    2.60     1718 138.0       190        12.2          4 D-penicil
## 2      NA    2.94     1612   6.2       183        11.2          4 D-penicil
## 3     200    2.74      918 147.3       302        11.5          4   placebo
## 4     178    3.00      746 178.3       119        12.0          4   placebo
## 5      NA    2.94      836 189.1        98        11.4          4   placebo
## 6      NA    3.02      650 124.0        99        11.7          4   placebo
##        age    sex     years event
## 1 58.76684 female 1.0951703     2
## 2 58.76684 female 1.0951703     2
## 3 70.56182 female 0.1396342     2
## 4 51.47027   male 1.5113350     2
## 5 51.47027   male 1.5113350     2
## 6 51.47027   male 1.5113350     2
```

For the illustration, 4 time-dependent predictors (`serBilir`, `SGOT`, `albumin` and `alkaline`) and 3 predictors measured at enrollment (`sex`, `age` and `drug`) were considered. We aim to predict the death without transplantation on patients suffering from primary biliary cholangitis (PBC) using clinical and socio-demographic predictors, considering the transplantation as a competing event.

4.2 Data management

To begin, we split the subjects into two datasets: (i) one dataset to train the random forest using 2/3 of patients; (ii) one dataset to predict on the other 1/3 of patients. The random seed is set to 1234 for replication purpose.

```
set.seed(1234)
id <- unique(pbc2$id)
id_sample <- sample(id, length(id)*2/3)
id_row <- which(pbc2$id %in% id_sample)
pbc2_train <- pbc2[id_row,]
pbc2_pred <- pbc2[-id_row,]
```

Then, we build the dataframe `timeData_train` in the longitudinal format (i.e., one observation per row) for the longitudinal predictors including: `id` the unique patient identifier; `time` the observed time measurements; `serBilir`, `SGOT`, `albumin` and `alkaline` the longitudinal predictors. We also build the dataframe `fixedData_train` with the time-fixed predictors including: `id` the unique patient identifier; `age`, `drug` and `sex` predictors measured at enrollment. The nature of each predictor needs to be properly defined with `as.factor()` function for categorical predictors (e.g., `drug` and `sex`).

```
timeData_train <- pbc2_train[,c("id","time",
                                "serBilir","SGOT",
                                "albumin","alkaline")]
fixedData_train <- unique(pbc2_train[,c("id","age","drug","sex")])
```

4.3 Specification of the models for the time-dependent predictors

The first step of the random forest building consists in specify the mixed model of each longitudinal predictor through a list containing the fixed and random formula for the fixed effect and random effects of the mixed models, respectively. Here, we assume a linear trajectory for `serBilir`, `albumin` and `alkaline`, and quadratic trajectory for `SGOT`. Although, we restricted this example to linear and quadratic functions of time, we note that any function can be considered including splines.

```
timeVarModel <- list(serBilir = list(fixed = serBilir ~ time,
                                     random = ~ time),
                     SGOT = list(fixed = SGOT ~ time + I(time^2),
                                 random = ~ time + I(time^2)),
                     albumin = list(fixed = albumin ~ time,
                                    random = ~ time),
                     alkaline = list(fixed = alkaline ~ time,
                                     random = ~ time))
```

For this illustration, the outcome object contains a list with `type` set to `surv` (for survival data) and `Y` contain's a dataframe in wide format (one subject per row) with: `id` the unique patient identifier; `years` the time-to-event data; `event` the event indicator.

```
Y <- list(type = "surv",
          Y = unique(pbc2_train[,c("id","years","event")]))
```

4.4 Random forest building

We build the random forest using `DynForest()` function with the following code:

```
res_dyn <- DynForest(timeData = timeData_train,
                     fixedData = fixedData_train,
                     timeVar = "time", idVar = "id",
                     timeVarModel = timeVarModel, Y = Y,
                     ntree = 200, mtry = 3, nodesize = 2, minsplit = 3,
                     cause = 2, ncores = 3, seed = 1234)
```

In a survival context with multiple events, it is necessary to specify the event of interest with the argument `cause`. We thus fixed `cause` $= 2$ to specify the event of interest (i.e., the death event). For the hyperparameters, we arbitrarily chose `mtry` $= 3$, `nodesize` $= 2$ and `minsplit` $= 3$ and we will discuss this point in Section 4.8.

Overall information about the random forest can be output with the `summary()` function as displayed below for our example:

```
summary(res_dyn)
```

```
## DynForest executed for survival (competing risk) outcome
##  Splitting rule: Fine & Gray statistic test
##  Out-of-bag error type: Integrated Brier Score
```

```
##  Leaf statistic: Cumulative incidence function
## ----------------
## Input
##  Number of subjects: 208
##  Longitudinal: 4 predictor(s)
##  Numeric: 1 predictor(s)
##  Factor: 2 predictor(s)
## ----------------
## Tuning parameters
##  mtry: 3
##  nodesize: 2
##  minsplit: 3
##  ntree: 200
## ----------------
## ----------------
## DynForest summary
##  Average depth per tree: 6.61
##  Average number of leaves per tree: 28.04
##  Average number of subjects per leaf: 4.7
##  Average number of events of interest per leaf: 1.91
## ----------------
## Out-of-bag error based on Integrated Brier Score
##  Out-of-bag error: Not computed!
## ----------------
## Computation time
##  Number of cores used:
##  Time difference of 19.64107 mins
## ----------------
```

We executed `DynForest()` function for a survival outcome with competing events. In this mode, we use the Fine & Gray statistic test as the splitting rule and the cumulative incidence function (CIF) as the leaf statistic. To build the random forest, we included 208 subjects with 4 longitudinal (`Longitudinal`), 1 continuous (`Numeric`) and 2 categorical (`Factor`) predictors. The `summary()` function returns some statistics about the trees. For instance, we have on average 4.7 subjects and 1.9 death events per leaf. The number of subjects per leaf should always be higher than `nodesize` hyperparameter. OOB error should be first computed using `compute_OOBerror()` function (see Section 4.5) to be displayed on summary output.

To further investigate the tree structure, the split details can be output with the following code (for tree 1):

```
head(res_dyn$rf[,1]$V_split)
```

```
##           type id_node var_split feature   threshold   N Nevent depth
## 1 Longitudinal       1         3       1 -0.21993638 129     49     1
## 2 Longitudinal       2         2       1  5.59590783  26     21     2
## 3      Numeric       3         1      NA 61.83057715 103     28     2
## 4 Longitudinal       4         2       3  1.49826008  18     13     3
## 5       Factor       5         1      NA          NA   8      8     3
## 6 Longitudinal       6         3       2 -0.01010312  92     22     3
```

```
tail(res_dyn$rf[,1]$V_split)
```

```
##            type id_node var_split feature threshold N Nevent depth
## 48         Leaf     192        NA      NA        NA 4      2     8
## 49         Leaf     193        NA      NA        NA 2      2     8
## 50         Leaf     194        NA      NA        NA 2      1     8
## 51 Longitudinal     195         4       1 -27.58024 4      3     8
## 52         Leaf     390        NA      NA        NA 2      1     9
## 53         Leaf     391        NA      NA        NA 2      2     9
```

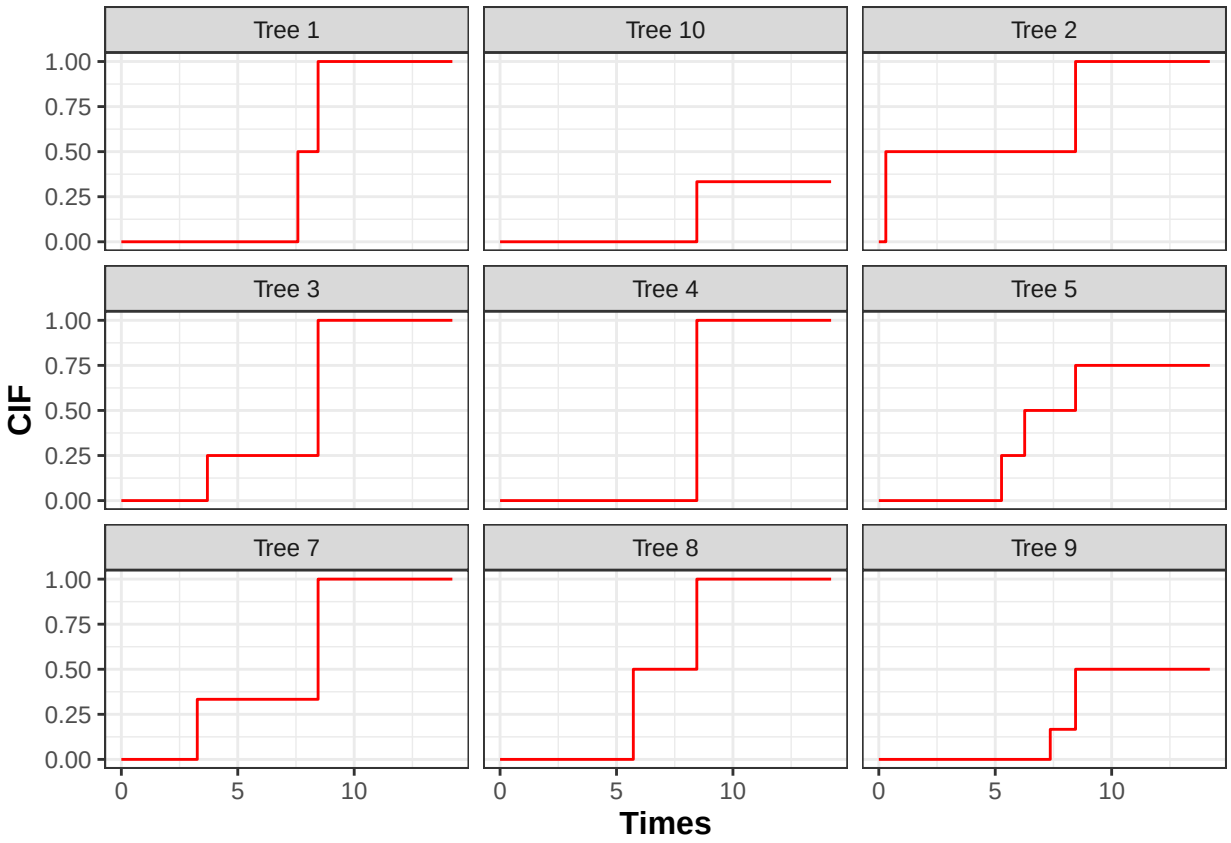

Figure 2: Estimated cumulative incidence functions of death before transplantation for subject 104 over 9 trees.

Looking at the head of `V_split` dataframe, we see that subjects were split at node 1 (`id_node`) using the first random-effect (`feature` = 1) of the third `Longitudinal` predictor (`var_split` = 3) with `threshold` = -0.2199. `var_split` = 3 corresponds to `albumin`, so subjects at node 1 with `albumin` values below to -0.2199 are assigned in node 2, otherwise in node 3. The last rows of random forest given by the tail of `V_split` provide the leaves descriptions. For instance, row 48, 4 subjects are included in leaf 192, and among 2 subjects have the event of interest.

Estimated cumulative incidence function (CIF) which in each leaf of a tree can be displayed using `$Y_pred` element of `$rf`. For instance, the CIF of the cause of interest for leaf 192 can be displayed using the following code:

```
plot(res_dyn$rf[,1]$Y_pred[[192]]$`2`, type = "l", col = "red",
     xlab = "Years", ylab = "CIF", ylim = c(0,1))
```

CIF of a single tree is not meant to be interpreted alone. The CIF should be average over all trees of the random forest. For a subject, estimated CIF over the random forest is obtained by averaging all the tree-specific CIF of the tree-leaf where the subject belongs. For instance, for subject 104, we display in Figure 2 the tree-specific CIF for the 9 first trees where this subject is used to grow the trees. This figure shows how the estimated CIF can be differ across the trees and requires to be averaged as each is calculated from information of the few subjects belonging to a leaf.

4.5 Out-of-bag error

The out-of-bag error (OOB) aims at assessing the prediction abilities of the random forest. With a survival outcome, the OOB error is evaluated using the integrated Brier score (IBS) (Gerds and Schumacher 2006). It

is computed using `compute_OOBerror()` function with an object of class `DynForest` as main argument, such as:

```
res_dyn_OOB <- compute_OOBerror(DynForest_obj = res_dyn,
                                ncores = 3)
```

`compute_OOBerror()` returns the OOB errors by individual (`$oob.err`). The overall OOB error for the random forest is obtained by averaging the individual specific OOB error.

```
mean(res_dyn_OOB$oob.err)
```

```
## [1] 0.1242997
```

We obtain an IBS of 0.124 computed from time 0 to the maximum event time. The time range can be modified using `IBS.min` and `IBS.max` arguments to define the minimum and maximum, respectively. To maximize the prediction ability of the random forest, the hyperparameters can be tuned, that is chosen as those that minimize the OOB error (see Section 4.8).

4.6 Prediction of the outcome

The `predict()` function allows to predict the outcome for a new subject using the trained random forest. The function requires the individual data: time-dependent predictors in `timeData` and time-fixed predictors in `fixedData`. For a survival outcome, dynamic predictions can be computed by fixing a prediction time (called landmark time, argument `t0`) from which prediction is made. In this case, only the history of the individual up to this landmark time (including the longitudinal and time-fixed predictors) will be used.

For the illustration, we only select the subjects still at risk at the landmark time of 4 years. We build the dataframe for those subjects and we predict the individual-specific CIF using `predict()` function as follows:

```
id_pred <- unique(pbc2_pred$id[which(pbc2_pred$years>4)])
pbc2_pred_tLM <- pbc2_pred[which(pbc2_pred$id %in% id_pred),]
timeData_pred <- pbc2_pred_tLM[,c("id","time",
                                  "serBilir","SGOT",
                                  "albumin","alkaline")]
fixedData_pred <- unique(pbc2_pred_tLM[,c("id","age","drug","sex")])
pred_dyn <- predict(object = res_dyn,
                    timeData = timeData_pred,
                    fixedData = fixedData_pred,
                    idVar = "id", timeVar = "time",
                    t0 = 4)
```

The `predict()` function provides several elements as described in Section 3.2. In addition, the `plot_CIF()` function can be used to display the CIF of the outcome (here death before transplantation) for subjects indicated with argument `id`. For instance, we compute the CIF for subjects 102 and 260 with the following code and display them in Figure 3.

```
plot_CIF(DynForestPred_obj = pred_dyn,
         id = c(102, 260))
```

The first year after the landmark time (at 4 years), we observe a rapid increase of the risk of death for subject 260 compared to subject 102. We also notice that after 10 years from landmark time, subject 260 has a probability of death almost three times higher that the one of subject 102.

4.7 Predictiveness variables

4.7.1 Variable importance

The main objective of the random forest is to predict an outcome. But usually, we are interested in identifying which predictors are the most predictive. The VIMP statistic (Hemant Ishwaran et al. 2008) can be computed using `compute_VIMP()` function. This function returns the VIMP statistic for each predictor with

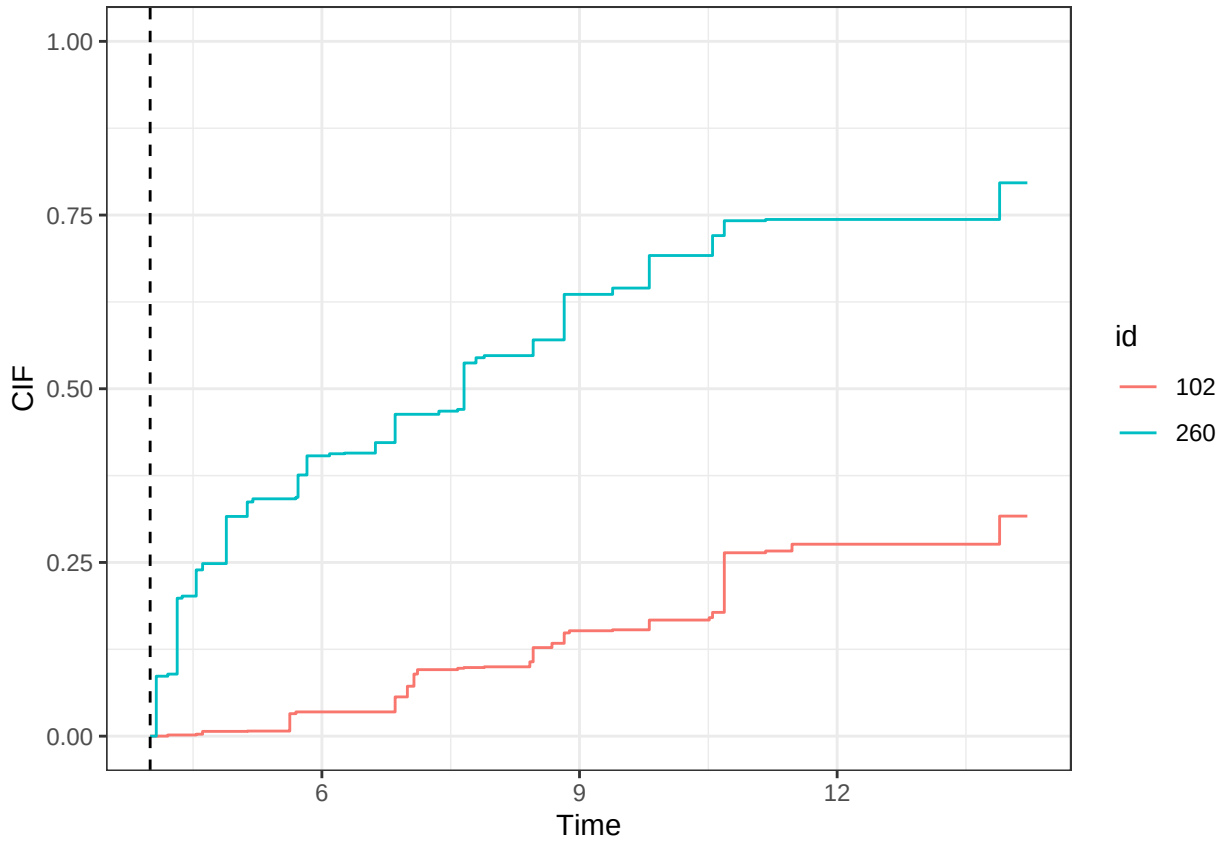

Figure 3: Predicted cumulative incidence function for subjects 102 and 260 from landmark time of 4 years (represented by the dashed vertical line).

`$Importance` element. These results can also be displayed using `plot()` function, either in absolute value by default or in percentage with `PCT` argument set to `TRUE`.

```
res_dyn_VIMP <- compute_VIMP(DynForest_obj = res_dyn,
                             ncores = 3, seed = 123)
plot(x = res_dyn_VIMP, PCT = TRUE)
```

The VIMP results are displayed in Figure 4A. The most predictive variables are **serBilir**, **albumin** and **age** with the largest VIMP percentage. By removing the association between **serBilir** and the event, the OOB error was increased by 28%.

In the case of correlated predictors, the predictors can be regrouped into dimensions and the VIMP can be computed at the dimension group level with the gVIMP statistic. Permutation is done for each variable of the group simultaneously. The gVIMP is computed with the `compute_gVIMP()` function in which the `group` argument defines the group of predictors as a list. For instance, with two groups of predictors (named group1 and group2), the gVIMP statistic is computed using the following code:

```
group <- list(group1 = c("serBilir","SGOT"),
              group2 = c("albumin","alkaline"))
res_dyn_gVIMP <- compute_gVIMP(DynForest_obj = res_dyn,
                               group = group, ncores = 3,
                               seed = 123)
plot(x = res_dyn_gVIMP, PCT = TRUE)
```

Similar to VIMP statistic, the gVIMP results can be displayed using `plot()` function. The Figure 4B shows that group1 has the highest gVIMP percentage with 32%.

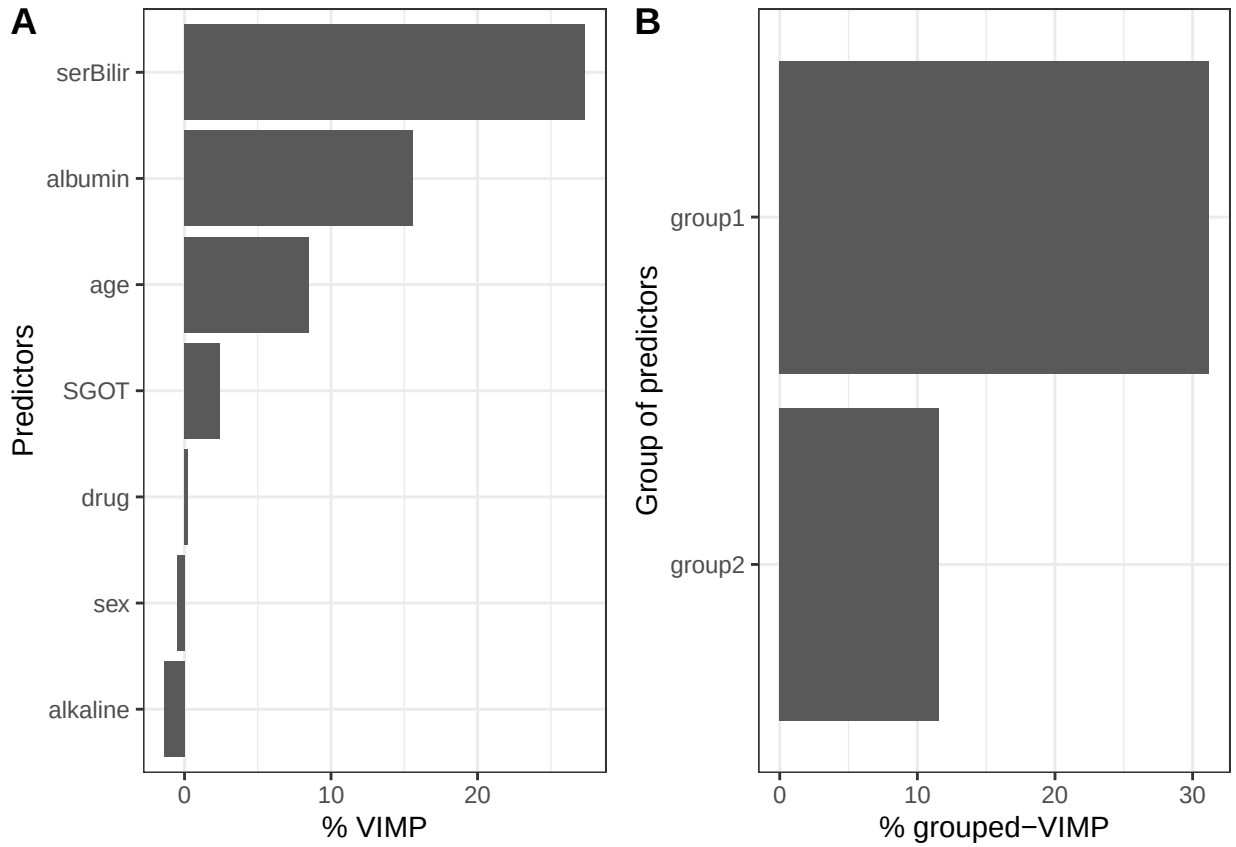

Figure 4: (A) VIMP statistic and (B) grouped-VIMP statistic displayed as a percentage of loss in OOB error of prediction. `group1` includes `serBilir` and `SGOT`; `group2` includes `albumin` and `alkaline`.

To compute the gVIMP statistic, the groups can be defined regardless of the number of predictors. However, the comparison between the groups may be harder when group sizes are very different.

4.7.2 Minimal depth

To go further into the understanding of the tree building process, the `var_depth()` function extracts information about the average minimal depth by feature (`$min_depth`), the minimal depth for each feature and each tree (`$var_node_depth`), the number of times that the feature is used for splitting for each feature and each tree (`$var_count`).

Using an object from `var_depth()` function, `plot()` function allows to plot the distribution of the average minimal depth across the trees. `plot_level` argument defines how the average minimal depth is plotted, by predictor or feature.

The distribution of the minimal depth level is displayed in Figure 5 by predictor and feature. Note that the minimal depth level should always be interpreted with the number of trees where the predictor/feature is found. Indeed, to accurately appreciate the importance of a variable minimal depth, the variable has to be systematically part of the candidates at each node. This is why we strongly advice to compute the minimal depth on random forest with `mtry` hyperparameter chosen at its maximum (as done below).

```
res_dyn_max <- DynForest(timeData = timeData_train,
                         fixedData = fixedData_train,
                         timeVar = "time", idVar = "id",
                         timeVarModel = timeVarModel, Y = Y,
                         ntree = 200, mtry = 7, nodesize = 2, minsplit = 3,
                         cause = 2, ncores = 3, seed = 1234)
```

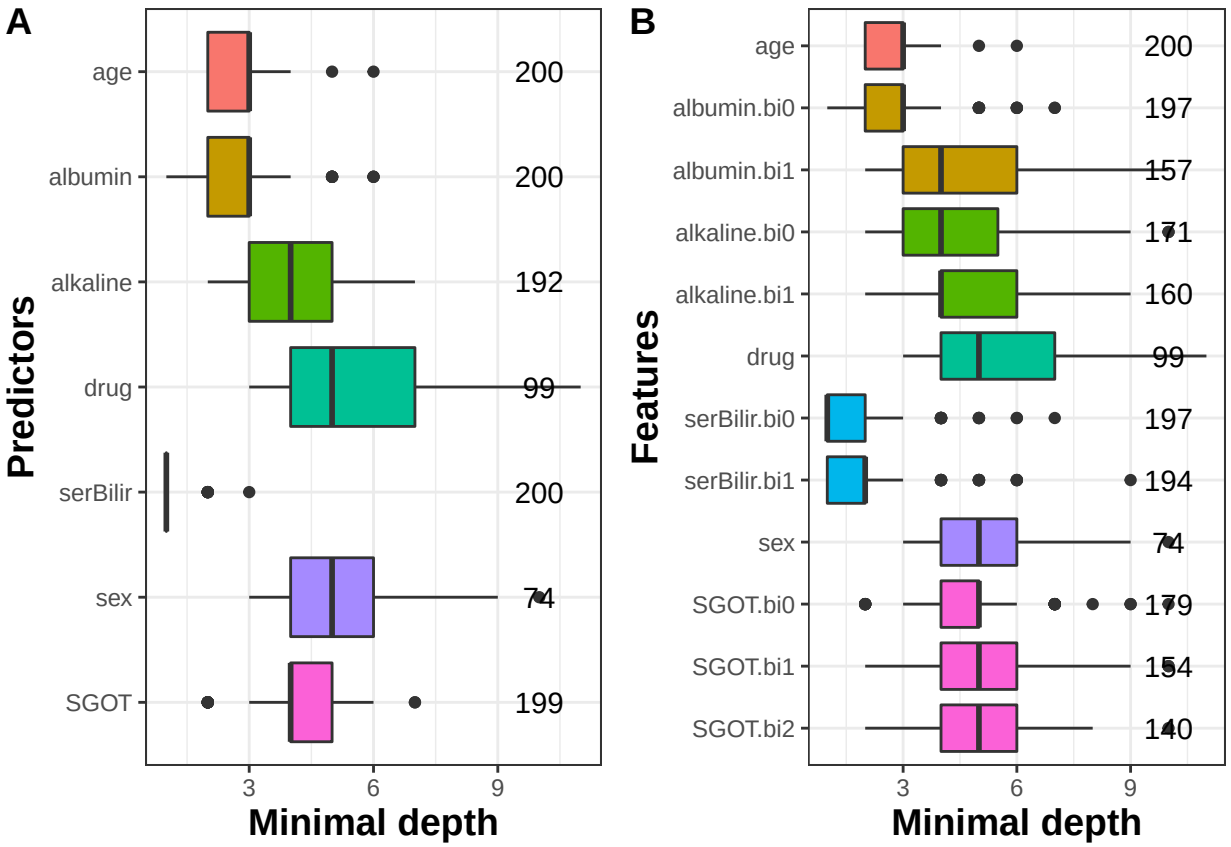

Figure 5: Average minimal depth level by predictor (A) and feature (B).

```
depth_dyn <- var_depth(DynForest_obj = res_dyn_max)
plot(x = depth_dyn, plot_level = "predictor")
plot(x = depth_dyn, plot_level = "feature")
```

In our example, we ran a random forest with mtry hyperparameter set to its maximum (i.e., `mtry` = 7) and we computed the minimal depth on this random forest. We observe that `serBilir`, `albumin` and `age` have the lowest minimal depth, indicating these predictors are used to split the subjects at early stages in 200 out of 200 trees, i.e., 100% (Figure 5A). The minimal depth level by feature (Figure 5B) provides more advanced details about the tree building process. For instance, we can see that the random-effects of `serBilir` (indicating by bi0 and bi1 in the graph) are the earliest features used on 197 and 194 out of 200 trees, respectively.

4.8 Guidelines to tune the hyperparameters

The predictive performance of the random forest strongly depends on the hyperparameters `mtry`, `nodesize` and `minsplit`. They should therefore be chosen thoroughly. `nodesize` and `minsplit` hyperparameters control the tree depth. The trees need to be deep enough to ensure that the predictions are accurate. By default, we fixed `nodesize` and `minsplit` at the minimum, that is `nodesize` = 1 and `minsplit` = 2. However, with a large number of individuals, the tree depth could be slightly decreased by increasing these hyperparameters in order to reduce the computation time.

`mtry` hyperparameter defines the number of predictors randomly drawn at each node. By default, we chose `mtry` equal to the square root of the number of predictors as usually recommended (Bernard, Heutte, and Adam 2009). However, this hyperparameter should be carefully tuned with possible values between 1 and the number of predictors. Indeed, the predictive performance of the random forest is highly related to this hyperparameter.

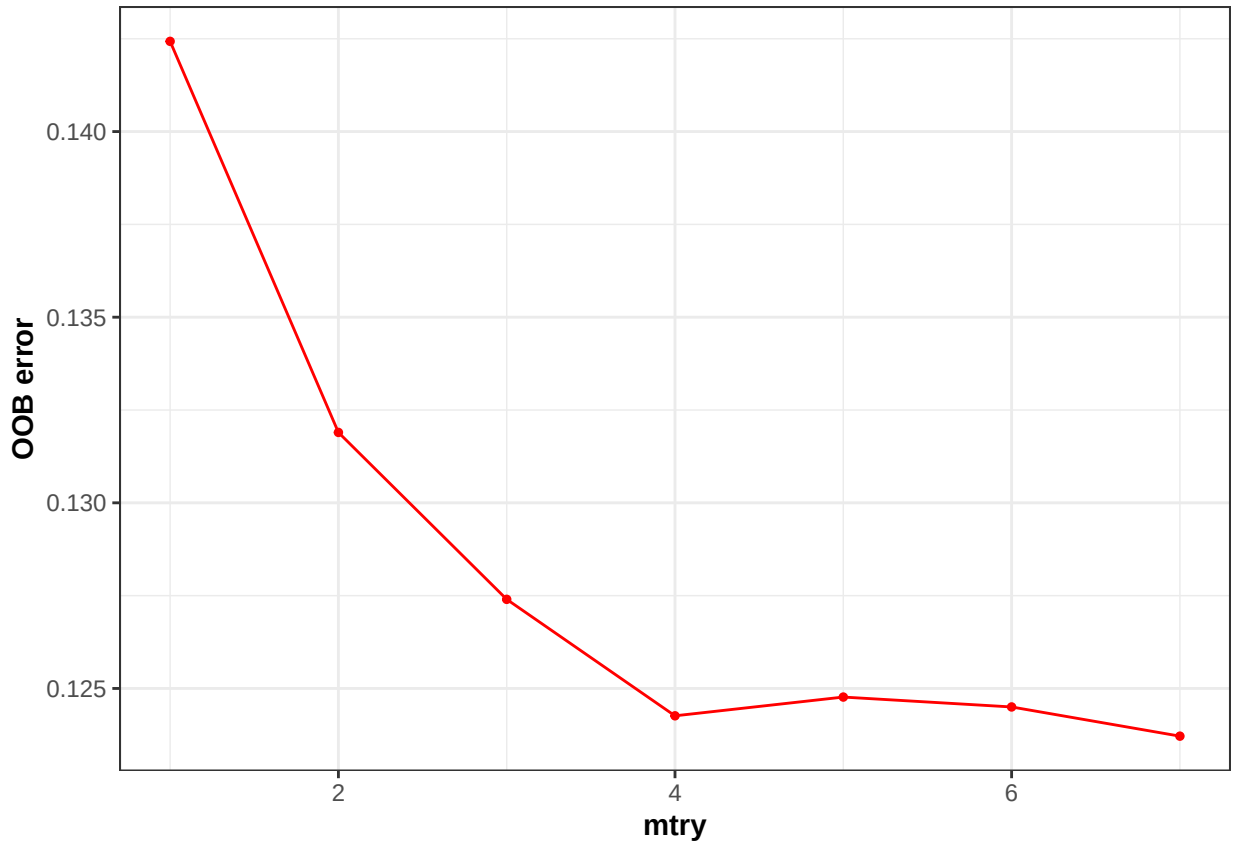

Figure 6: OOB error according to `mtry` hyperparameter. The optimal value was found for the maximum value `mtry` = 7.

In the illustration, we tuned `mtry` for every possible values (1 to 7). Figure 6 displays the OOB error according to `mtry` hyperparameter.

We can see on this figure large OOB error difference according to `mtry` hyperparameter. In particular, we observe the worst predictive performance for lower values, then similar results with values from 4 to 7. The optimal value (i.e., with the lowest OOB error) was found with the maximum value `mtry` = 7. This graph reflects how crucial it is to carefully tune this hyperparameter.

5 How to use DynForest R package with a categorical outcome?

In this section, we use **DynForest** in a classification perspective using `pbc2` data. For the illustration purpose, we want to predict the death between 4 and 10 years on subjects still at risk at 4 years from the repeated data up to 4 years. Note that this is only for illustrative purpose as this technique does not handle censoring correctly.

5.1 Data management

For the illustration, we select patients still at risk at 4 years and we recode the `event` variable with `event` = 1 for subjects who died during between 4 years and 10 years, whereas subjects with transplantation were recoded `event` = 0, as the subjects still alive. We split the subjects into two datasets: (i) one dataset to train the random forest using 2/3 of patients; (ii) one dataset to predict on the other 1/3 of patients.

```
library("DynForest")
pbc2 <- pbc2[which(pbc2$years>4&pbc2$time<=4),]
pbc2$event <- ifelse(pbc2$event==2, 1, 0)
```

```
pbc2$event[which(pbc2$years>10)] <- 0
set.seed(1234)
id <- unique(pbc2$id)
id_sample <- sample(id, length(id)*2/3)
id_row <- which(pbc2$id%in%id_sample)
pbc2_train <- pbc2[id_row,]
pbc2_pred <- pbc2[-id_row,]
```

We use the same strategy as in the survival context (Section 4) to build the random forest, with the same predictors and the same association for time-dependent predictors.

```
timeData_train <- pbc2_train[,c("id","time",
                                "serBilir","SGOT",
                                "albumin","alkaline")]
timeVarModel <- list(serBilir = list(fixed = serBilir ~ time,
                                     random = ~ time),
                     SGOT = list(fixed = SGOT ~ time + I(time^2),
                                 random = ~ time + I(time^2)),
                     albumin = list(fixed = albumin ~ time,
                                    random = ~ time),
                     alkaline = list(fixed = alkaline ~ time,
                                     random = ~ time))
fixedData_train <- unique(pbc2_train[,c("id","age","drug","sex")])
```

With a categorical outcome, the definition of the output object is slightly different. We should specify `type`="factor" to define the outcome as categorical, and the dataframe in `Y` should contain only 2 columns, the variable identifier `id` and the outcome `event`.

```
Y <- list(type = "factor",
          Y = unique(pbc2_train[,c("id","event")]))
```

5.2 The random forest building

We executed `DynForest()` function to build the random forest with hyperparameters `mtry` = 7 and `nodesize` = 2 as follows:

```
res_dyn <- DynForest(timeData = timeData_train,
                     fixedData = fixedData_train,
                     timeVar = "time", idVar = "id",
                     timeVarModel = timeVarModel,
                     mtry = 7, nodesize = 2,
                     Y = Y, ncores = 3, seed = 1234)
```

5.3 Out-of-bag error

With a categorical outcome, the OOB error is evaluated using a missclassification criterion. This criterion can be computed with `compute_OOBerror()` function and the results of the random forest can be displayed using `summary()`:

```
res_dyn_OOB <- compute_OOBerror(DynForest_obj = res_dyn,
                                ncores = 3)
```

```
summary(res_dyn_OOB)
```

```
## DynForest executed for categorical outcome
##  Splitting rule: Minimize weighted within-group Shannon entropy
##  Out-of-bag error type: Missclassification
##  Leaf statistic: Majority vote
```

```
## ----------------
## Input
##  Number of subjects: 150
##  Longitudinal: 4 predictor(s)
##  Numeric: 1 predictor(s)
##  Factor: 2 predictor(s)
## ----------------
## Tuning parameters
##  mtry: 7
##  nodesize: 2
##  ntree: 200
## ----------------
## ----------------
## DynForest summary
##  Average depth per tree: 5.84
##  Average number of leaves per tree: 16.66
##  Average number of subjects per leaf: 9.26
## ----------------
## Out-of-bag error based on Missclassification
##  Out-of-bag error: 0.2333
## ----------------
## Computation time
##  Number of cores used:
##  Time difference of 14.4759 mins
## ----------------
```

In this illustration, we built the random forest using 150 subjects because we only kept the subjects still at risk at landmark time at 4 years and split the dataset in 2/3 for training and 1/3 for testing. We have on average 9.3 subjects per leaf, and the average depth level per tree is 5.8. This random forest predicted the wrong outcome for 23% of the subjects. The random forest performances can be optimized by choosing the `mtry` and `nodesize` hyperparameters that minimized the OOB missclassification rate.

5.4 Prediction of the outcome

We can predict the probability of death between 4 and 10 years on subjects still at risk at landmark time at 4 years. In classification mode, the predictions are performed using the majority vote. The prediction over the trees is thus a category of the outcome along with the proportion of the trees that lead to this category. Predictions are computed using `predict()` function, then a dataframe can be easily built from the returning object to get the prediction and probability of the outcome for each subject:

```
timeData_pred <- pbc2_pred[,c("id","time",
                              "serBilir","SGOT",
                              "albumin","alkaline")]
fixedData_pred <- unique(pbc2_pred[,c("id","age","drug","sex")])
pred_dyn <- predict(object = res_dyn,
                    timeData = timeData_pred,
                    fixedData = fixedData_pred,
                    idVar = "id", timeVar = "time",
                    t0 = 4)
```

```
head(data.frame(pred = pred_dyn$pred_indiv,
                proba = pred_dyn$pred_indiv_proba))
```

```
##     pred proba
## 101    0 0.965
## 104    0 0.785
## 106    1 0.560
## 108    0 0.955
## 112    1 0.515
## 114    0 0.640
```

As shown in this example, some predictions are made with varying confidence from 51.5% for subject 112 to 96.5% for subject 101. We predict for instance no event for subject 101 with a probability of 96.5% and an event for subject 106 with a probability of 56.0%.

5.5 Predictiveness variables

5.5.1 Variable importance

The most predictive variables can be identified using `compute_VIMP()` and displayed using `plot()` function as follows:

```
res_dyn_VIMP <- compute_VIMP(DynForest_obj = res_dyn_OOB,
                             ncores = 3, seed = 123)
plot(x = res_dyn_VIMP, PCT = TRUE)
```

Again, we found that the most predictive variable is `serBilir`: when perturbating `serBilir`, the OOB error was increased by 23%.

5.5.2 Minimal depth

The minimal depth is computed using `var_depth()` function and is displayed at predictor and feature levels using `plot()` function. The results are displayed in Figure 7 using the random forest with maximal `mtry` hyperparameter value (i.e., `mtry` = 7) for a better understanding.

```
depth_dyn <- var_depth(DynForest_obj = res_dyn_OOB)
plot(x = depth_dyn, plot_level = "predictor")
plot(x = depth_dyn, plot_level = "feature")
```

We observe that `serBilir` and `albumin` have the lowest minimal depth and are used to split the subjects in almost all the trees (198 and 194 out of 200 trees, respectively) (Figure 7A). Figure 7B provides further results. In particular, this graph shows that the random intercept (indicated by bi0) of `serBilir` and `albumin` are the earliest predictors used to split the subjects and are present in 195 and 189 out of 200 trees, respectively.

6 How to use DynForest R package with a continuous outcome?

In this section, we present an illustration of **DynForest** with a continuous outcome. **DynForest** was used on a simulated dataset with 200 subjects and 10 predictors (6 time-dependent and 4 time-fixed predictors). The 6 longitudinal predictors were generated using a linear mixed model with linear trajectory according to time. We considered 6 measurements by subject (at baseline and then randomly drawn around theoretical annual visits up to 5 years). Then, we generated the continuous outcome using a linear regression with the random intercept of marker 1 and random slope of marker 2 as linear predictors. We generated two datasets (`data_simu1` and `data_simu2`), one for each step (training and prediction). These datasets are available in the **DynForest** package.

The aim of this illustration is to predict the continuous outcome using time-dependent and time-fixed predictors.

6.1 Data management

First of all, we load the data and we build the mandatory objects needed to execute `DynForest()` function that are `timeData_train` for time-dependent predictors and `fixedData_train` for time-fixed predictors. We specify the model for the longitudinal predictors in `timeVarModel` object. We considered linear trajectories over time for the 6 longitudinal predictors.

```
timeData_train <- data_simu1[,c("id","time",
                                paste0("marker",seq(6)))]
timeVarModel <- lapply(paste0("marker",seq(6)),
                       FUN = function(x){
                         fixed <- reformulate(termlabels = "time",
                                              response = x)
```

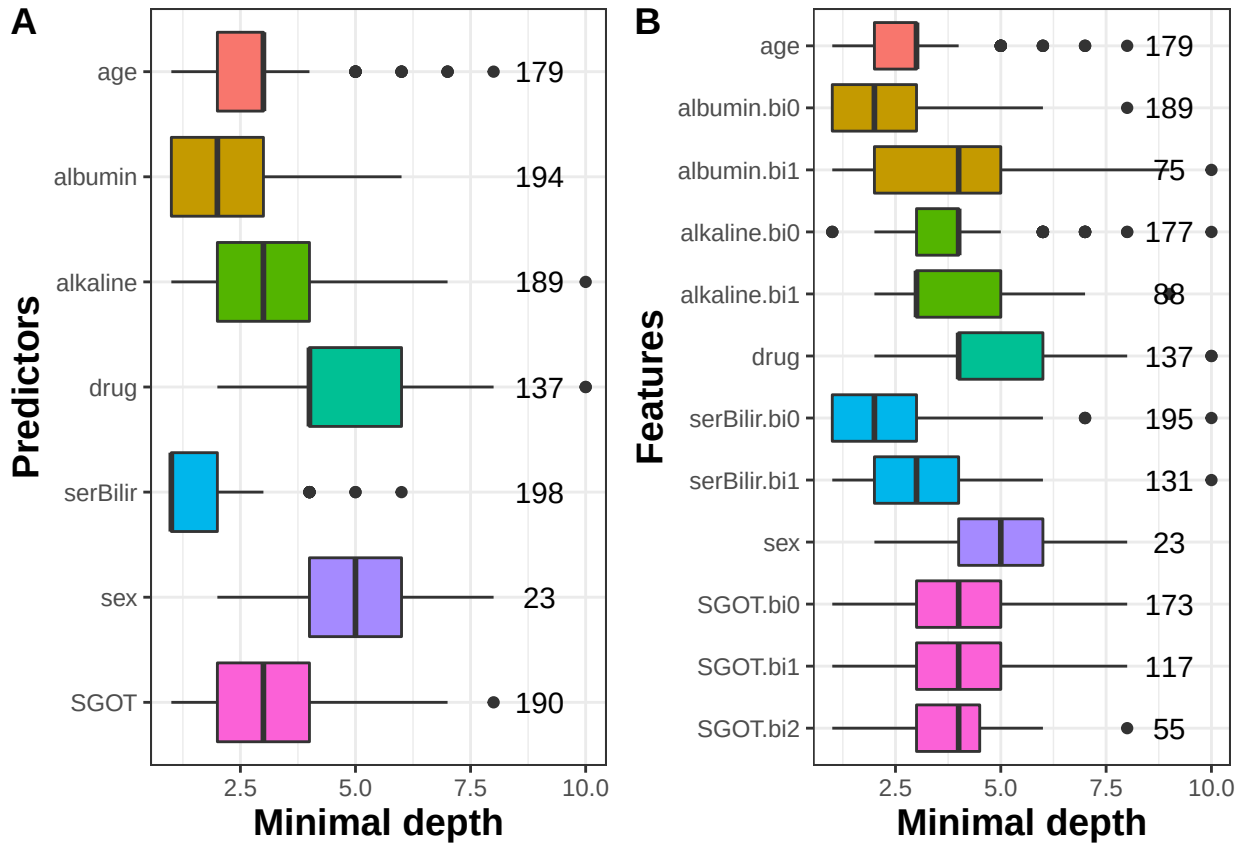

Figure 7: Average minimal depth by predictor (A) and feature (B).

```
                          random <- ~ time
                          return(list(fixed = fixed, random = random))
                        })
fixedData_train <- unique(data_simu1[,c("id",
                                        "cont_covar1","cont_covar2",
                                        "bin_covar1","bin_covar2")])
```

To define the `Y` object for a continuous outcome, the `type` argument should be set to `numeric` to run the random forest in regression mode. The dataframe `Y` should include two columns with the unique identifier `id` and the continuous outcome, here `Y_res`.

```
Y <- list(type = "numeric",
          Y = unique(data_simu1[,c("id","Y_res")]))
```

6.2 The random forest building

To build the random forest, we chose default hyperparameters (i.e., `ntree` = 200 and `nodesize` = 1), except for `mtry` which was fixed at its maximum (i.e., `mtry` = 10). We ran `DynForest()` function with the following code:

```
res_dyn <- DynForest(timeData = timeData_train,
                     fixedData = fixedData_train,
                     timeVar = "time", idVar = "id",
                     timeVarModel = timeVarModel,
```

```
                      mtry = 10, Y = Y,
                      ncores = 3, seed = 1234)
```

6.3 Out-of-bag error

For continuous outcome, the OOB error is evaluated using the mean square error (MSE). We used `compute_OOBerror()` function to compute the OOB error and we provided overall results with `summary()` function as shown below:

```
res_dyn_OOB <- compute_OOBerror(DynForest_obj = res_dyn,
                                ncores = 3)
```

```
summary(res_dyn_OOB)
```

```
## DynForest executed for continuous outcome
##  Splitting rule: Minimize weighted within-group variance
##  Out-of-bag error type: Mean square error
##  Leaf statistic: Mean
## ----------------
## Input
##  Number of subjects: 200
##  Longitudinal: 6 predictor(s)
##  Numeric: 2 predictor(s)
##  Factor: 2 predictor(s)
## ----------------
## Tuning parameters
##  mtry: 10
##  nodesize: 1
##  ntree: 200
## ----------------
## ----------------
## DynForest summary
##  Average depth per tree: 9.06
##  Average number of leaves per tree: 126.47
##  Average number of subjects per leaf: 3.03
## ----------------
## Out-of-bag error based on Mean square error
##  Out-of-bag error: 4.3626
## ----------------
## Computation time
##  Number of cores used:
##  Time difference of 47.41606 mins
## ----------------
```

The random forest was executed in regression mode (for a continuous outcome). The splitting rule aimed to minimize the weighted within-group variance. We built the random forest using 200 subjects and 10 predictors (6 time-dependent and 4 time-fixed predictors) with hyperparameters `ntree` = 200, `mtry` = 10 and `nodesize` = 1. As we can see, `nodesize` = 1 leads to deeper trees (the average depth by tree is 9.1) and few subjects by leaf (3 on average). We obtained 4.4 for the MSE. This quantity can be minimized by tuning hyperparameters `mtry` and `nodesize`.

6.4 Prediction of the outcome

In regression mode, the tree and leaf-specific means are averaged across the trees to get a unique prediction over the random forest. `predict()` function provides the individual predictions. We first created the `timeData` and `fixedData` from the testing sample `data_simu2`. We then predicted the continuous outcome by running `predict()` function:

```
timeData_pred <- data_simu2[,c("id","time",
                               paste0("marker",seq(6)))]
fixedData_pred <- unique(data_simu2[,c("id","cont_covar1","cont_covar2",
                                       "bin_covar1","bin_covar2")])
pred_dyn <- predict(object = res_dyn,
                    timeData = timeData_pred,
                    fixedData = fixedData_pred,
                    idVar = "id", timeVar = "time")
```

Individual predictions can be extracted using the following code:

```
head(pred_dyn$pred_indiv)
```

```
##         1         2         3         4         5         6
##  5.218430 -1.286623  0.871417  1.531534  5.295980  7.910508
```

For instance, we predicted 5.22 for subject 1, -1.29 for subject 2 and 0.87 for subject 3.

6.5 Predictiveness variables

In this illustration, we want to evaluate if **DynForest** can identify the true predictors (i.e., random intercept of marker1 and random slope of marker2). To do this, we used the minimal depth which allows to understand the random forest at the feature level.

Minimal depth information can be extracted using `var_depth()` function and can be displayed with `plot()` function. For the purpose of this illustration, we displayed the minimal depth in Figure 8 by predictor and by feature.

```
depth_dyn <- var_depth(DynForest_obj = res_dyn)
plot(x = depth_dyn, plot_level = "predictor")
plot(x = depth_dyn, plot_level = "feature")
```

We observe in Figure 8A that marker2 and marker1 have the lowest minimal depth, as expected. To go further, we also looked into the minimal depth computed on features. We perfectly identified the random slope of marker2 (i.e., marker2.bi1) and the random intercept of marker1 (i.e., marker1.bi0) as the predictors in this simulated dataset.

7 Discussion

The **DynForest R** package provides an easy-to-use random forests methodology for predictors that may contain longitudinal variables possibly measured irregularly with error. Note that the method can also be used without any longitudinal predictors such as other random forests packages.

We implemented several statistics to identify the predictive ability of each variable with the VIMP, gVIMP and average minimal depth. For survival outcome, compared to **randomForestSRC** R package, we considered two different stopping criteria `nodesize` and `minsplit` to favor the deepest forests possible and avoid suboptimal splits. We designed **DynForest** to be as user-friendly as possible. To achieve that, we implemented various functions to summarize and display the results, and provided a step-by-step analysis in the three modes; survival, categorical and continuous.

Nevertheless, several improvements could be considered in the future. We used linear mixed models for longitudinal continuous outcomes but alternatives strategies could be considered such as PACE algorithm (Yao, Müller, and Wang 2005) based on functional data analysis. We could also consider different natures of longitudinal predictors (e.g., binary) for which generalized linear mixed models could be used. **DynForest** currently handles continuous, categorical and survival (with possibly competing events) outcomes. But other outcomes could be envisaged such as curves, recurrent events or interval-censored time-to-events. We leave these perspectives for future releases.

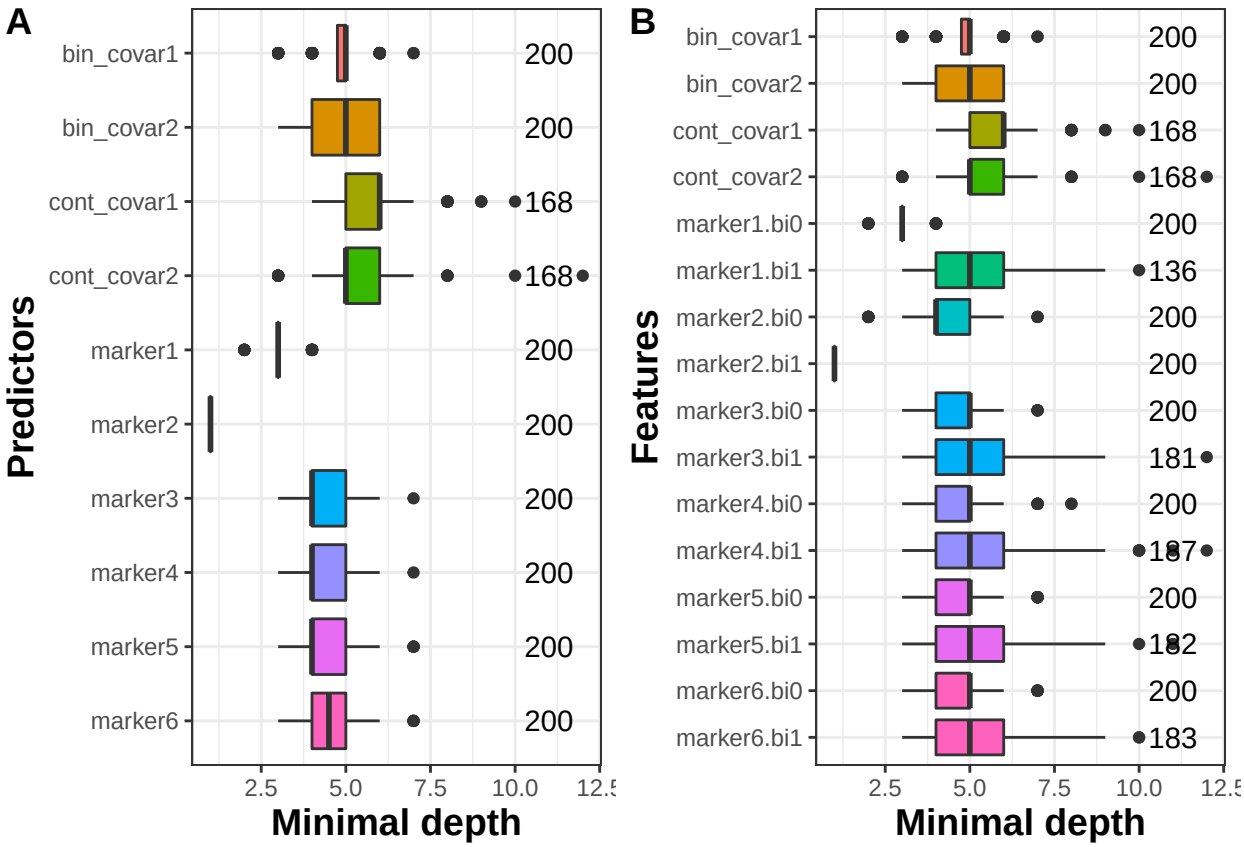

Figure 8: Average minimal depth level by predictor (A) and by feature (B).

Computational details

The results in this paper were obtained using **R** 4.2.0 with the **DynForest** 1.1.0 package. **R** itself and all packages used are available from the Comprehensive **R** Archive Network (CRAN) at `https://CRAN.R-project.org/`.

Acknowledgments

We thank Dr. Louis Capitaine for **FrechForest R** code used in **DynForest**.

This work was funded by the French National Research Agency (ANR-18-CE36-0004-01 for project DyMES), and the French government in the framework of the PIA3 ("Investment for the future") (project reference 17-EURE-0019) and in the framework of the University of Bordeaux's IdEx "Investments for the Future" program / RRI PHDS.

References

Aalen, Odd. 1976. "Nonparametric Inference in Connection with Multiple Decrement Models." *Scandinavian Journal of Statistics* 3 (1): 15–27.

Aalen, Odd O., and Søren Johansen. 1978. "An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations." *Scandinavian Journal of Statistics* 5 (3): 141–50.

Bernard, Simon, Laurent Heutte, and Sébastien Adam. 2009. "Influence of Hyperparameters on Random Forest Accuracy." In *International Workshop on Multiple Classifier Systems*, 171–80. Springer.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32. `https://doi.org/10.1023/A:1010933404324`.

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–94.

Devaux, Anthony. 2022. *DynForest: Random Forest with Multivariate Longitudinal Predictors*. https://CRAN.R-project.org/package=DynForest.

Devaux, Anthony, Catherine Helmer, Carole Dufouil, Robin Genuer, and Cécile Proust-Lima. 2022. "Random Survival Forests for Competing Risks with Multivariate Longitudinal Endogenous Covariates." *arXiv Preprint arXiv:2208.05801*.

Gerds, Thomas A., and Martin Schumacher. 2006. "Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times." *Biometrical Journal* 48 (6): 1029–40. https://doi.org/10.1002/bimj.200610301.

Gray, Bob. 2020. *Cmprsk: Subdistribution Analysis of Competing Risks*. https://CRAN.R-project.org/package=cmprsk.

Gray, Robert J. 1988. "A Class of K-Sample Tests for Comparing the Cumulative Incidence of a Competing Risk." *The Annals of Statistics* 16 (3): 1141–54.

Gregorutti, Baptiste, Bertrand Michel, and Philippe Saint-Pierre. 2017. "Correlation and Variable Importance in Random Forests." *Statistics and Computing* 27 (3): 659–78.

Ishwaran, Hemant, Thomas A. Gerds, Udaya B. Kogalur, Richard D. Moore, Stephen J. Gange, and Bryan M. Lau. 2014. "Random Survival Forests for Competing Risks." *Biostatistics* 15 (4): 757–73. https://doi.org/10.1093/biostatistics/kxu010.

Ishwaran, Hemant, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. 2008. "Random Survival Forests." *The Annals of Applied Statistics* 2 (3): 841–60. https://doi.org/10.1214/08-AOAS169.

Ishwaran, H., and U. B. Kogalur. 2022. *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*. https://cran.r-project.org/package=randomForestSRC.

Laird, Nan M., and James H. Ware. 1982. "Random-Effects Models for Longitudinal Data." *Biometrics* 38 (4): 963–74. https://doi.org/10.2307/2529876.

Murtaugh, Paul A., E. Rolland Dickson, Gooitzen M. Van Dam, Michael Malinchoc, Patricia M. Grambsch, Alice L. Langworthy, and Chris H. Gips. 1994. "Primary Biliary Cirrhosis: Prediction of Short-Term Survival Based on Repeated Patient Visits." *Hepatology* 20 (1): 126–34. https://doi.org/10.1002/hep.1840200120.

Nelson, Wayne. 1969. "Hazard Plotting for Incomplete Failure Data." *Journal of Quality Technology* 1 (1): 27–52. https://doi.org/10.1080/00224065.1969.11980344.

Peto, Richard, and Julian Peto. 1972. "Asymptotically Efficient Rank Invariant Test Procedures." *Journal of the Royal Statistical Society: Series A (General)* 135 (2): 185–98.

Proust-Lima, Cécile, Viviane Philipps, and Benoit Liquet. 2017. "Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm." *Journal of Statistical Software* 78 (2): 1–56. https://doi.org/10.18637/jss.v078.i02.

R Core Team. 2019. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Sène, Mbéry, Jeremy MG Taylor, James J Dignam, Hélène Jacqmin-Gadda, and Cécile Proust-Lima. 2016. "Individualized Dynamic Prediction of Prostate Cancer Recurrence with and Without the Initiation of a Second Treatment: Development and Validation." *Statistical Methods in Medical Research* 25 (6): 2972–91.

Shannon, Claude Elwood. 1948. "A Mathematical Theory of Communication." *The Bell System Technical Journal* 27 (3): 379–423.

Therneau, Terry M. 2022. *A Package for Survival Analysis in r*. https://CRAN.R-project.org/package=survival.

Wright, Marvin N., and Andreas Ziegler. 2017. "Ranger: A Fast Implementation of Random Forests for High Dimensional Data in c++ and r." *Journal of Statistical Software* 77 (1). https://doi.org/10.18637/jss.v077.i01.

Yao, Fang, Hans-Georg Müller, and Jane-Ling Wang. 2005. "Functional Data Analysis for Sparse Longitudinal Data." *Journal of the American Statistical Association* 100 (470): 577–90. https://doi.org/10.1198/016214504000001745.

References