Random forests for time series

Benjamin GOEHRY, Hui YAN, Yannig GOUDE, Pascal MASSART, Jean-Michel POGGI

EDF Lab & Univ. Paris-Sud
1. Standard random forests

2. Adaptation to time series

3. Application to load forecasting & Conclusion
Standard random forests
We have some stationary data set $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$, $(X_i, Y_i) \in \mathbb{R}^p \times \mathbb{R}$ and

$$Y = f(X) + \epsilon$$

**Goal:** estimate the regression function $f$.

**How:** random forests
A partitioning of $[0, 1]^2$ and the associated binary tree.
Breiman’s random forest

Parameters: number of trees $M$, number of observations per tree $\alpha_n$, size of the random set of variables $m_{try}$

Repeat for each tree:

- Draw randomly $\alpha_n \leq n$ points among the $n$ points with or without replacement.
- Repeat recursively at each node:
  - choose a random set of $m_{try}$ variables among the $p$ variables and apply the CART criterion on this subset.
  - Cut on the best split.

Key step: bootstrapping

Randomly drawing $\alpha_n \leq n$ observations with replacement.

**Pros:** adapted to i.i.d observations.

**Cons:** destroys the underlying structure.

Example:

![Original load.](image1)

![Bootstraped load.](image2)
Adaptation to time series
Solution: Block bootstrap

Replace the standard bootstrap with a block bootstrap variant to subsample time series during the tree construction phase.

→ Dependence structure preserved.

Example:

Original load.  Bootstrapped load.  24h block bootstrapped
New algorithm

Parameters: number of trees $M$, number of observations per tree $\alpha_n$, size of the random set of variables $m_{\text{try}}$, block size $l_n$

Repeat for each tree:

- Draw $\alpha_n \leq n$ observations using a block bootstrap variant with parameter $l_n$.
- Repeat recursively at each node:
  - Chose a random set of $m_{\text{try}}$ variables among the $p$ variables and apply the CART criterion on this subset.
  - Cut on the best split.
Block bootstrap variants

Non-overlapping block bootstrap\(^1\)

\[ Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7, Y_8, Y_9 \]

\[ Y_1^*, Y_2^*, Y_3^*, Y_4^*, Y_5^*, Y_6^*, Y_7^*, Y_8^*, Y_9^* \]

Moving block bootstrap\(^2\)

\[ Y_1, Y_2, Y_3, Y_4, Y_5, Y_6, Y_7, Y_8, Y_9 \]

\[ Y_1^*, Y_2^*, Y_3^*, Y_4^*, Y_5^*, Y_6^*, Y_7^*, Y_8^*, Y_9^* \]

\(^1\)E. Carlstein. *The use of subseries values for estimating the variance of a general statistic from a stationary sequence.* 1986.


Existing packages for trees/RF: party, rpart, randomForest, ranger\(^3\), etc.

We propose an extension of ranger called **rangerts**.

**New code parameters:**

- **bootstrap.ts:** "circular", "moving", "non-overlap" (and others)
- **block.size:** number of consecutive observations per block
- **by.end:** build blocks by the end of the series or not
- **period:** seasonality period (only for seasonal variant)

**Code example:**

```r
forest_ts ← ranger(Y ~ ., data, bootstrap.ts = "moving", block.size = l_n)
forecast_ts ← predict(forest_ts, data_test)$prediction
```

Application to load forecasting & Conclusion
Dataset: load of a building called UnivLab Patrick. One observation per hour over one year. Access to the temperature and schedule. Training January-October, validation November, test December.

Goal: Load forecasting at a 24 hour horizon $Y_t$.

Predictor variables:
- $Y_{t-24}$ & $Y_{t-7\times24}$;
- $\text{Temp}_t$;
- $\text{Schedule}_t$;
- $\text{Hour}_t, \text{InstantWeek}_t, \text{DayType}_t, \text{Toy}_t$.

---

4C. Miller, F. Meggers. The building data genome project: An open, public data set from non-residential building electrical meters. 2017
Comparison to the standard random forest

Performance of the variants for $m_{\text{try}} = 2$.

Evolution of the performance for each variant according the block length $l_n$. 
Take-home message

- Introduced a new way to incorporate the dependence structure in random forests.
- Improve the performance over the standard random forests.
- Variable importance can also be redefined.
References


Random forests can be used to compute the variable importance.

**Mean Decrease Accuracy**: if a variable is not important, then permuting its value should not change prediction accuracy. The importance of the variable $X^{(j)}$ is defined by

$$VI\left( X^{(j)} \right) = \frac{1}{M} \sum_{m=1}^{M} \left( \frac{\text{errOOB}_{m}^{(j)}}{\text{errOOB}_{m}} \right)$$

New definition:

$$VI\left( X^{(j)} \right) = \frac{1}{M} \sum_{m=1}^{M} \left( \frac{\text{errOOB}_{m}^{(j)}}{\text{errOOB}_{m}} - \text{errOOB}_{m} \right)$$

---

Standard variable importance vs Block variable importance

Non-overlapping standard variable importance

Non-overlapping block variable importance with $l_n = 24$. 
**Software aspect in R (help)**

```r
ranger(formula = NULL, data = NULL, num.trees = 500, mtry = NULL,
importance = "none", write.forest = TRUE, probability = FALSE,
min.node.size = NULL, max.depth = NULL, replace = TRUE,
sample.fraction = ifelse(replace, 1, 0.632), case.weights = NULL,
class.weights = NULL, splitrule = NULL, num.random.splits = 1,
alpha = 0.5, minprop = 0.1, split.select.weights = NULL,
always.split.variables = NULL, respect.unordered.factors = NULL,
scale.permutation.importance = FALSE, keep.inbag = FALSE,
inbag = NULL, holdout = FALSE, quantreg = FALSE,
oob.error = TRUE, num.threads = NULL, save.memory = FALSE,
verbose = TRUE, seed = NULL, dependent.variable.name = NULL,
status.variable.name = NULL, classification = NULL,
bootstrap.ts = NULL, by.end = TRUE, block.size = 10, period = 1)
```

**New ranger function with all the parameters**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>bootstrap.ts</td>
<td>Bootstrapping mode: empty for iid observations, &quot;nonoverlapping&quot; is default, &quot;moving&quot; for moving blocks, &quot;circular&quot; for circular blocks, &quot;stationary&quot; for stationary blocks, and &quot;seasonal&quot; for seasonal blocks.</td>
</tr>
<tr>
<td>by.end</td>
<td>Logical. Build block by the end of time series or not. Default = TRUE.</td>
</tr>
<tr>
<td>block.size</td>
<td>Number of observations in one block only if bootstrap by block is activated (bootstrap.ts has non null value).</td>
</tr>
<tr>
<td>period</td>
<td>Number of steps of one period. Only for the 'seasonal' block bootstrap.</td>
</tr>
</tbody>
</table>

**The new parameters**