Random forest for functional data

Nathalie Vialaneix

nathalie.vialaneix@inrae.fr
http://www.nathalievialaneix.eu

European Courses in Advanced Statistics

ECAS-SFdS course on random forest
Fréjus (France), October 8-13, 2023

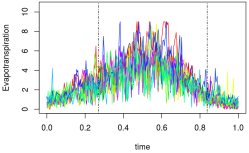
RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

INRAE

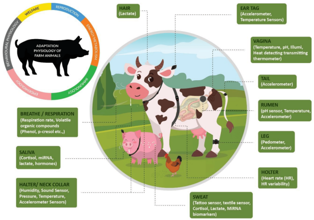
Functional data in a nutshell

Functional data... just functions (in mathematical sense) [**Ramsay and Silverman, 2005**, **Ramsay and Silverman, 2002**].

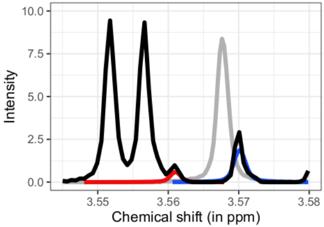
Examples: time series (mostly): weather, wearable sensors, chemiometrics spectra, ...



[**Picheny et al., 2019**]

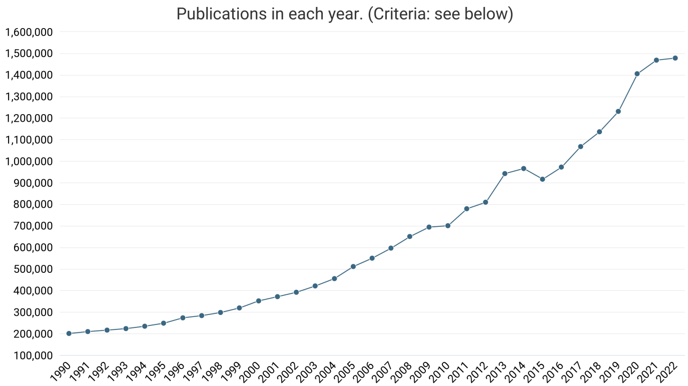


[**Neethirajan, 2020**]

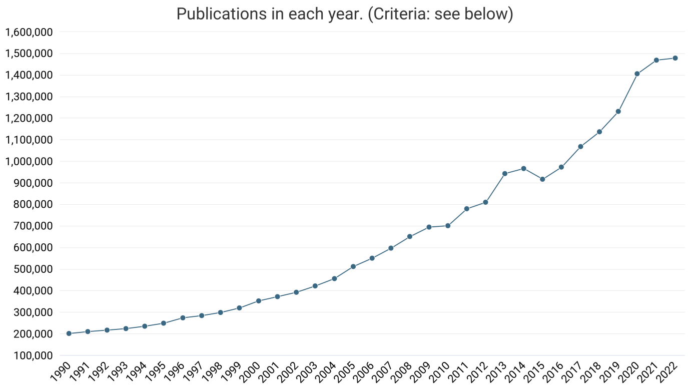


[**Lefort et al., 2021**]

Time series is the new trend?

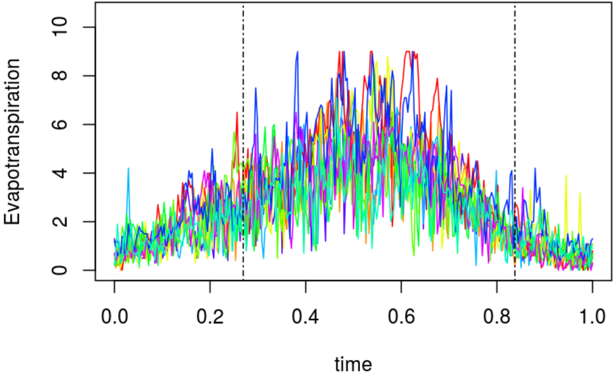


Time series is the new trend?



Disclaimer: Forecasting is a specific case, not covered by this class

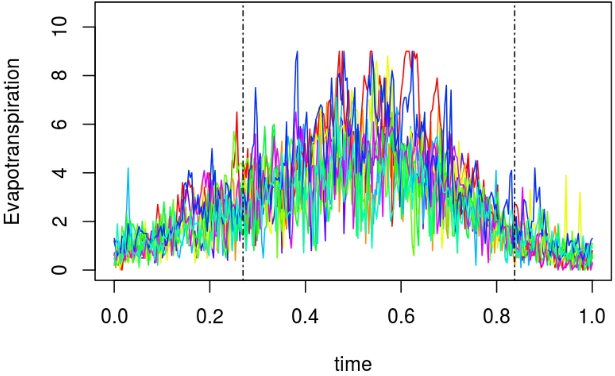
Scientific question



$\xrightarrow{?}$

Purpose: prediction of a target quantity (*e.g.*, yield) from functional data (*e.g.*, weather time series)

Scientific question



$\xrightarrow{?}$

Purpose: prediction of a target quantity (*e.g.*, yield) from functional data (*e.g.*, weather time series)

Here: ? = random forest

Extensions of random forest for time series

- Similarity based techniques
  - Proximity forest [**Lucas et al., 2019**] (restricted to classification)
  - Fréchet forest [**Capitaine et al., 2020**]

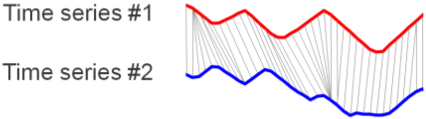


Image by courtesy of Charlotte Pelletier

Extensions of random forest for time series

- Similarity based techniques
  - Proximity forest [Lucas et al., 2019] (restricted to classification)
  - Fréchet forest [Capitaine et al., 2020]
- Interval based techniques
  - Time Series Forest [Deng et al., 2013] and its extension [Middlehurst et al., 2020]
  - RISE [Lines et al., 2018]

Extensions of random forest for time series

- Similarity based techniques
  - Proximity forest [**Lucas et al., 2019**] (restricted to classification)
  - Fréchet forest [**Capitaine et al., 2020**]
- Interval based techniques
  - Time Series Forest [**Deng et al., 2013**] and its extension [**Middlehurst et al., 2020**]
  - RISE [**Lines et al., 2018**]
- Dictionnary or symbolic representation based techniques:
  - TS-CHIEF [**Shifaz et al., 2020**] (combines all types of splits including dictionnary based splits based on work of [**Schäfer, 2015**])
  - (multivariate time series) symbolic representation of time series [**Baydogan and Runger, 2015**]

Similarity based techniques

Interval based techniques

Dictionnary or symbolic representation based techniques

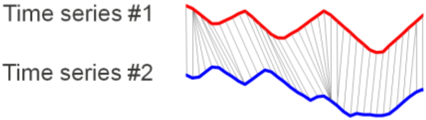
Improving interpretability: interval selection

You don't know what to do with your time series?

Use distances! (or kernels)

Numeric time series

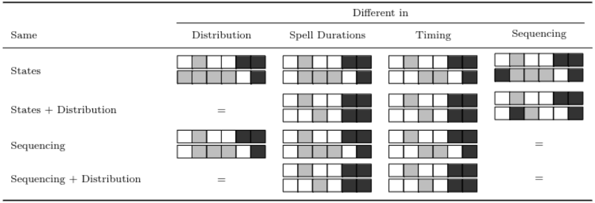
DTW [**Sakoe and Chiba, 1978**], Derivative DTW [**Keogh and Pazzani, 2001**], ...



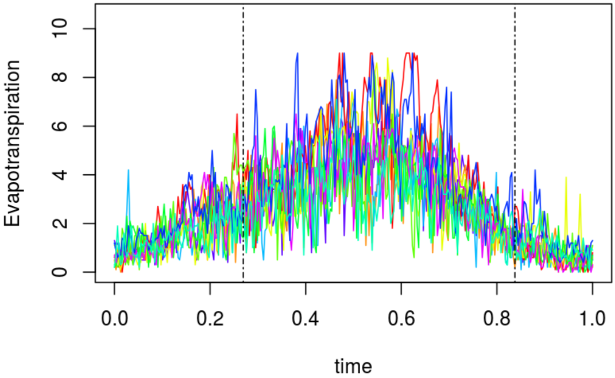
See: R package **TSclust**

Categorical time series

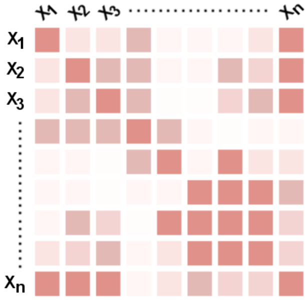
χ^2-metric, optimal matching, edit distances, ... [**Massoni et al., 2013**, **Studer and Ritschard, 2016**]



Use distances!

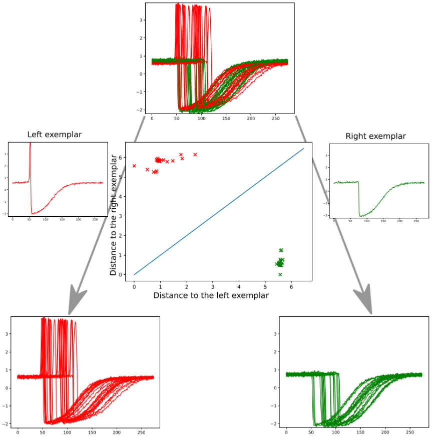


$\xrightarrow{?}$



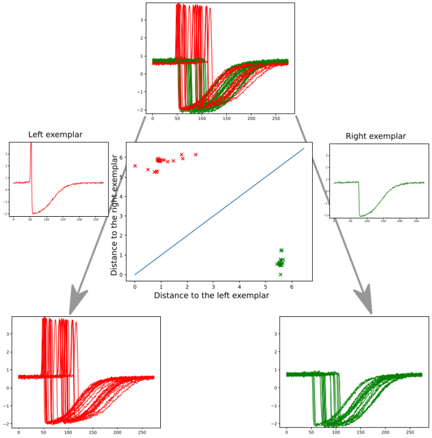
But now: I don't have variables anymore to define splits!

Proximity forest [Lucas et al., 2019] (classification only)



Splits defined by splitter pairs and distances.

Proximity forest [Lucas et al., 2019] (classification only)

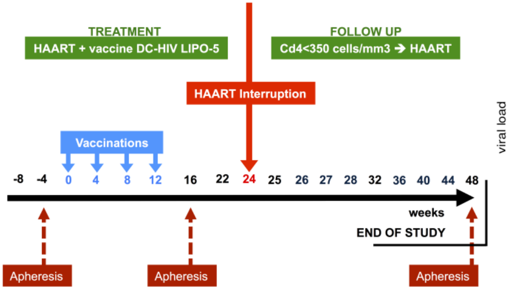


Splits defined by splitter pairs and distances.

In practice: select best "splitter" pair among R (5) randomly chosen splitter pairs using Gini.

Generalization: Fréchet forests [Capitaine et al., 2020]

Inputs: repeated time series



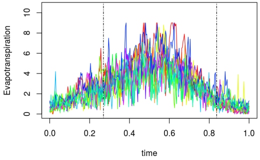
Example: $n = 17$ patients $\times$ $p = 5,398$ gene expression time series

Goal: predict viral load (also time series)

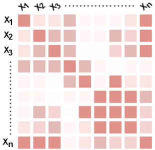
Notation: $X_i = (X_i^{(1)}, \ldots, X_i^{(p)})$ where $X_i^{(j)} \in (\mathcal{X}_j, d_j)$ (metric space)
$Y_i \in (\mathcal{Y}, d)$ (also a metric space)

some slides by courtesy of R. Genuer

Basics on Fréchet things... (similar to kernels)



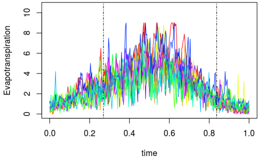
$\xrightarrow{?}$



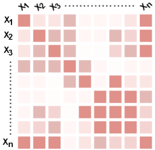
d_j

Can I compute distances and variance just using d_j?

Basics on Fréchet things... (similar to kernels)



$\xrightarrow{?}$



d_j

Can I compute distances and variance just using d_j?

Yes! [**Fréchet, 1906**, **Peterson and Müller, 2019**]:

- empirical Fréchet mean of $(X_i^{(j)})_{i \in \mathcal{C}}$:

$$\overline{X^{(j)}} \in \arg\min_{z \in (\mathcal{X}_j, d_j)} \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} d_j^2\left(X_i^{(j)}, z\right)$$

- empirical Fréchet variance:

$$\mathcal{V}_{\mathcal{C}} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} d_j^2\left(X_i^{(j)}, \overline{X^{(j)}}\right)$$

In short: Fréchet split for $X^{(j)}$

1. Fréchet 2-means **[Genolini et al., 2016]** $\rightarrow$ partition of $(X_i^{(j)})_{i \in \mathcal{C}}$ into $\mathcal{C}_L^j$ and $\mathcal{C}_R^j$

In short: Fréchet split for $X^{(j)}$

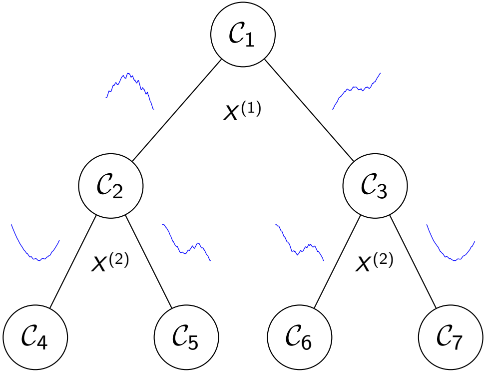
1. Fréchet 2-means **[Genolini et al., 2016]** $\rightarrow$ partition of $(X_i^{(j)})_{i \in \mathcal{C}}$ into $\mathcal{C}_L^j$ and $\mathcal{C}_R^j$

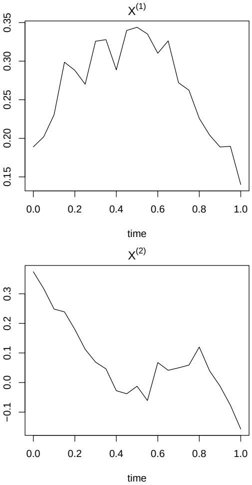
2. Quality of split:

$$\Phi^{(j)}(\mathcal{C}) - \left(\frac{|\mathcal{C}_L^j|}{|\mathcal{C}|} \Phi(\mathcal{C}_L^j) + \frac{|\mathcal{C}_R^j|}{|\mathcal{C}|} \Phi(\mathcal{C}_R^j) \right)$$

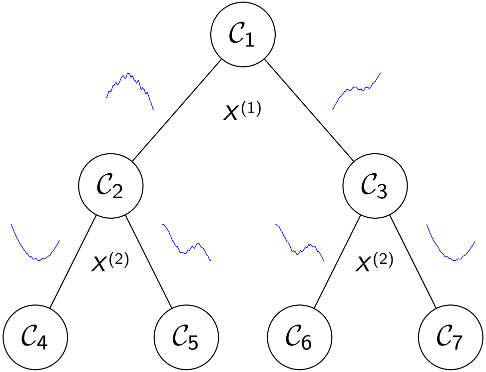
with Φ: Fréchet variance of Y

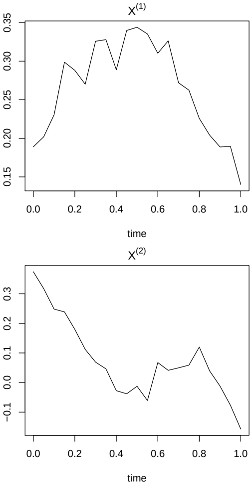
In short: Fréchet tree





In short: Fréchet tree





prediction: Fréchet mean of Y in $\mathcal{C}_5$

Summary: Fréchet random forests

https://github.com/Lcapitaine/FrechForest/tree/master (R package)

$$
\begin{tikzcd}
 & & \mathcal{L}_n \arrow[dll, "\text{Bootstrap}"'] \arrow[d] \arrow[drr] & & \\
\mathcal{L}_n^{\Theta_1} \arrow[d, "\text{RI Fréchet tree}"'] & \cdots & \mathcal{L}_n^{\Theta_\ell} \arrow[d] & \cdots & \mathcal{L}_n^{\Theta_q} \arrow[d] \\
\widehat{h}(.,\Theta_1,\Theta'_1) \arrow[drr, "\text{Fréchet mean}"'] & \cdots & \widehat{h}(.,\Theta_\ell,\Theta'_\ell) \arrow[d] & \cdots & \widehat{h}(.,\Theta_q,\Theta'_q) \arrow[dll] \\
 & & \widehat{h}_{\mathrm{FRF}}(.,\Theta_{1:q},\Theta'_{1:q}) & &
\end{tikzcd}
$$

Take home message of similarity based RF

- Proximity forest
  - splits based on random draw of two functions (X)
  - nodes based on distances between functions
  - restricted to classification

- Fréchet forest
  - splits and nodes based on distance based 2-means
  - more suited for multivariate function inputs
  - adapted to any type of outputs

Similarity based techniques

Interval based techniques

Dictionnary or symbolic representation based techniques

Improving interpretability: interval selection

Time Series Forest

Basic principles:

1. for a given tree: random sampling of intervals
2. for a given tree: compute summaries (mean, sd, slope for **[Deng et al., 2013]**)
3. define splits as usual based on these summaries

Implemented in Python package **ptys** (contains also most transformations or preprocessings described in this class).

Time Series Forest

Bootstrap

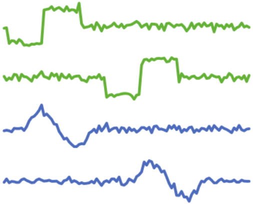
select intervals + compute summaries

averaging

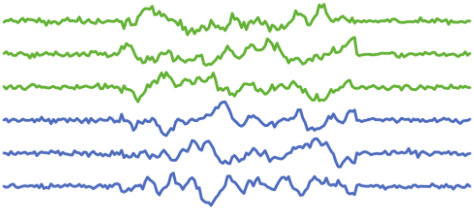
$$\begin{tikzcd}
 & & \mathcal{L}_n \arrow[dll] \arrow[d] \arrow[drr] & & \\
\mathcal{L}_n^{\Theta_1} \arrow[d] & \cdots & \mathcal{L}_n^{\Theta_\ell} \arrow[d] & \cdots & \mathcal{L}_n^{\Theta_q} \arrow[d] \\
\widehat{h}(., \Theta_1, \Theta'_1) \arrow[drr] & \cdots & \widehat{h}(., \Theta_\ell, \Theta'_\ell) \arrow[d] & \cdots & \widehat{h}(., \Theta_q, \Theta'_q) \arrow[dll] \\
 & & \widehat{h}_{\text{FRF}}(., \Theta_{1:q}, \Theta'_{1:q}) & &
\end{tikzcd}$$

More on summaries for time-series

- add more summaries: catch22 [**Middlehurst et al., 2020**])
- use basis decomposition, power spectrum (Fourier) or auto-correlation features [**Lines et al., 2018**] (HIVE-COTE)



$X(t) = \sum_{j=1}^{p} \beta_j \phi_j(t)$
(ϕ_j: basis functions)
$\longrightarrow (\beta_j)_j \subset \mathbb{R}^p$



$X_i \longrightarrow \text{FFT} \in \mathbb{R}^T$

Random Interval Spectral Ensemble (RISE) [Lines et al., 2018]

$$
\begin{tikzcd}
 & & \mathcal{L}_n \arrow[dll, "\text{random interval selection}"'] \arrow[d] \arrow[drr] & & \\
\mathcal{L}_n^{\mathcal{I}_1} \arrow[d, "\text{tree based on PS or ACT } (\sim p = 8)"'] & \cdots\cdots & \mathcal{L}_n^{\mathcal{I}_\ell} \arrow[d] & \cdots\cdots & \mathcal{L}_n^{\mathcal{I}_q} \arrow[d] \\
\widehat{h}(., \mathcal{I}_1, \Theta'_1) \arrow[drr, "\text{averaging}"'] & \cdots\cdots & \widehat{h}(., \mathcal{I}_\ell, \Theta'_\ell) \arrow[d] & \cdots\cdots & \widehat{h}(., \mathcal{I}_q, \Theta'_q) \arrow[dll] \\
 & & \widehat{h}_{\text{FRF}}(., \mathcal{I}_{1:q}, \Theta'_{1:q}) & &
\end{tikzcd}
$$

Similarity based techniques

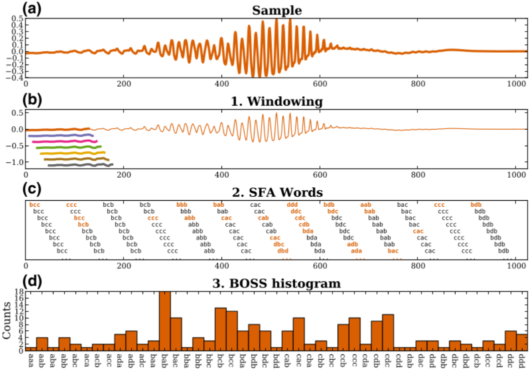
Interval based techniques

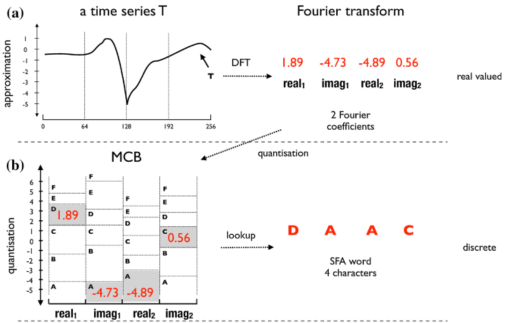
Dictionnary or symbolic representation based techniques

Improving interpretability: interval selection

Symbolic representations based on FFT and windowing

BOSS [Schäfer, 2015]



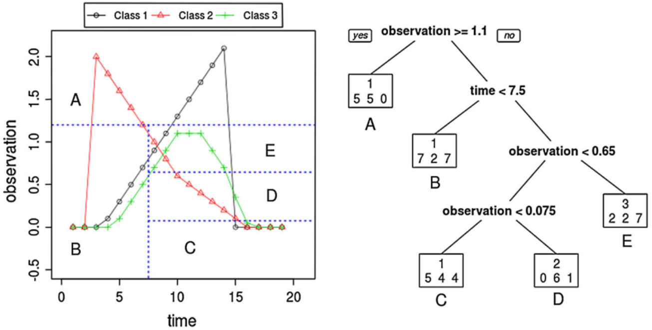


Based on: Fourier transform then symbolic representation.

A Java implementation exists.

Symbolic representation based on trees

[**Baydogan and Runger, 2015**]



Recode $X_i(t_k)$ using the tree partitionning (A, B, C, D, E) = proportion of the time series in each class.
$X_i \longrightarrow \mathbb{R}^{\sum_{t=1}^T N_t}$ (N_t: number of partitions induced by tree t).

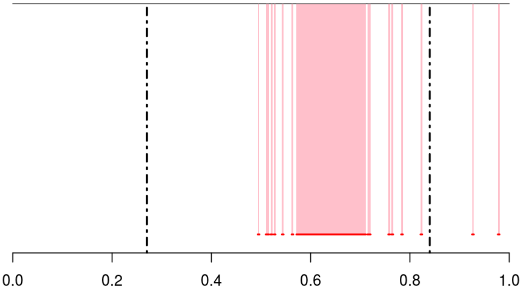
Similarity based techniques

Interval based techniques

Dictionnary or symbolic representation based techniques

Improving interpretability: interval selection

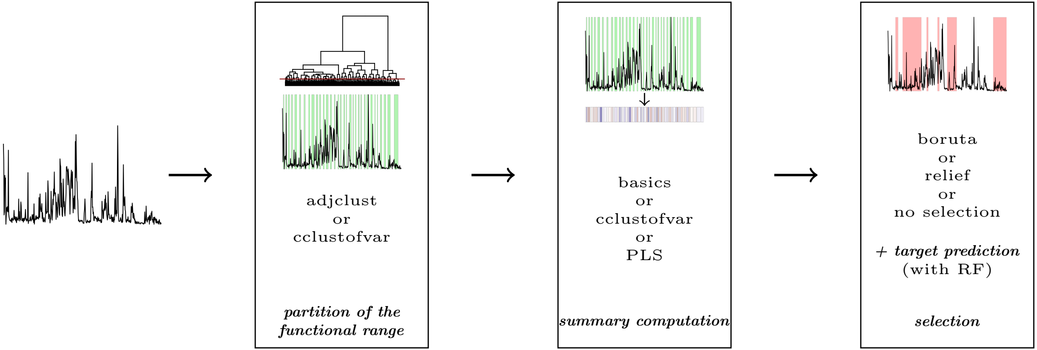
Scientific question



Purpose: Improve interpretability by selecting the most predictive intervals.

Overview of SFCB (Selection Forest for funCtion Based predictions)

[Servien and Vialaneix, 2023], R package **SISIR**



A focus on partition of the functional range

Two unsupervised and data-driven methods:

- Constrained hierarchical clustering **[Randriamihamison et al., 2021]** as in R package **adjclust** **[Ambroise et al., 2019]**: correlations between time steps + kernel based HC

A focus on partition of the functional range

Two unsupervised and data-driven methods:

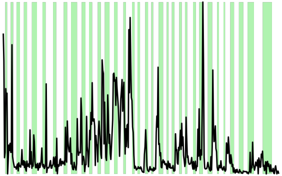
- Constrained hierarchical clustering [Randriamihamison et al., 2021] as in R package **adjclust** [Ambroise et al., 2019]: correlations between time steps + kernel based HC
- Clustering of variables (**ClustOfVar**, [Chavent et al., 2012]; also hierarchical and also based on correlation but using PCA-like criterion) but constrained to contiguity (implementation available in **SISIR**)

A focus on partition of the functional range

Two unsupervised and data-driven methods:

- Constrained hierarchical clustering **[Randriamihamison et al., 2021]** as in R package **adjclust** **[Ambroise et al., 2019]**: correlations between time steps + kernel based HC
- Clustering of variables (**ClustOfVar**, **[Chavent et al., 2012]**; also hierarchical and also based on correlation but using PCA-like criterion) but constrained to contiguity (implementation available in **SISIR**)

Output: dendrogram + cut $\Rightarrow$ intervals (or hierarchy or intervals)



A focus on summary of intervals

Three methods:

- Unsupervised: mean and sd

A focus on summary of intervals

Three methods:

- Unsupervised: mean and sd
- Supervised: 1st PLS component (same idea in [Poterie et al., 2019] for group-based RF; or other authors . . .)

A focus on summary of intervals

Three methods:

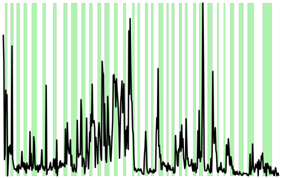
- Unsupervised: mean and sd
- Supervised: 1st PLS component (same idea in **[Poterie et al., 2019]** for group-based RF; or other authors ...)
- for cClustOfVar only: composite variable obtained from **ClustOfVar** (similar to PC)

A focus on summary of intervals

Three methods:

- Unsupervised: mean and sd
- Supervised: 1st PLS component (same idea in **[Poterie et al., 2019]** for group-based RF; or other authors . . .)
- for cClustOfVar only: composite variable obtained from **ClustOfVar** (similar to PC)

Output: $\mathbb{R}^K$ vector



$\longrightarrow$

A focus on variable selection

Two methods:

- RF based variable selection: Boruta [Kursa and Rudnicki, 2010] (other methods available like the excellent **VSURF** [Genuer et al., 2015], see [Speiser et al., 2019, Degenhardt et al., 2019])

  I am so eager to know more!!

A focus on variable selection

Two methods:

- RF based variable selection: Boruta [Kursa and Rudnicki, 2010] (other methods available like the excellent **VSURF** [Genuer et al., 2015], see [Speiser et al., 2019, Degenhardt et al., 2019])

  I am so eager to know more!!

- a standard VS method: Relief (extended to regression [Kira and Rendell, 1992, Robnik-Šikonja and Kononenko, 1997])

  I am so eager to know more!!

A focus on variable selection

Two methods:

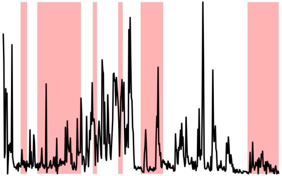
- RF based variable selection: Boruta [**Kursa and Rudnicki, 2010**] (other methods available like the excellent **VSURF** [**Genuer et al., 2015**], see [**Speiser et al., 2019**, **Degenhardt et al., 2019**])

  I am so eager to know more!!

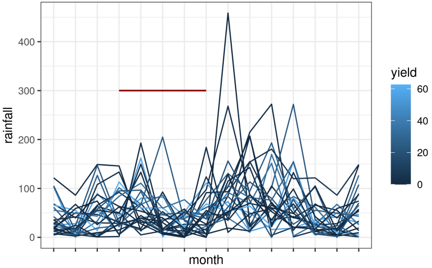
- a standard VS method: Relief (extended to regression [**Kira and Rendell, 1992**, **Robnik-Šikonja and Kononenko, 1997**])

  I am so eager to know more!!

Output: selected summaries corresponding to intervals



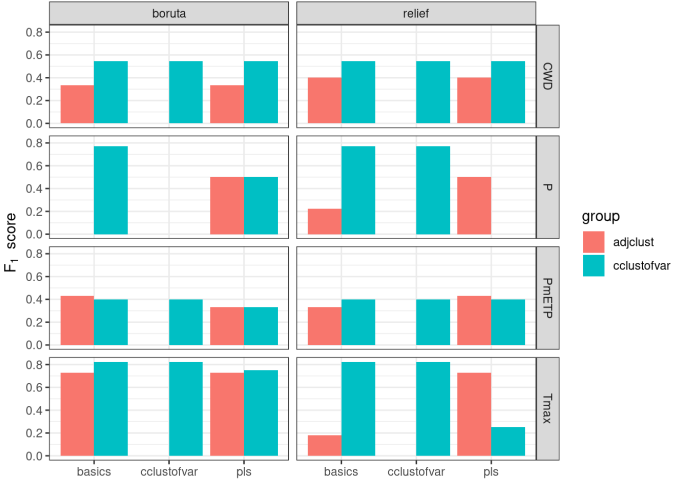
Application to black truffle & weather time series



Dataset courtesy from authors of [Baragatti et al., 2019] and fully available at https://doi.org/10.57745/KMH2GP.

- X: $p = 15$ monthly measures (4: rainfall, sun, …) from January of year N to March of year $N+1$ for $N \in [\![1925, 1949]\!]$ ($n = 25$)
- Y: yield of truffles year $N+1$
- expert knowledge of important periods for each weather measurement

A brief overview of the comparison between variants

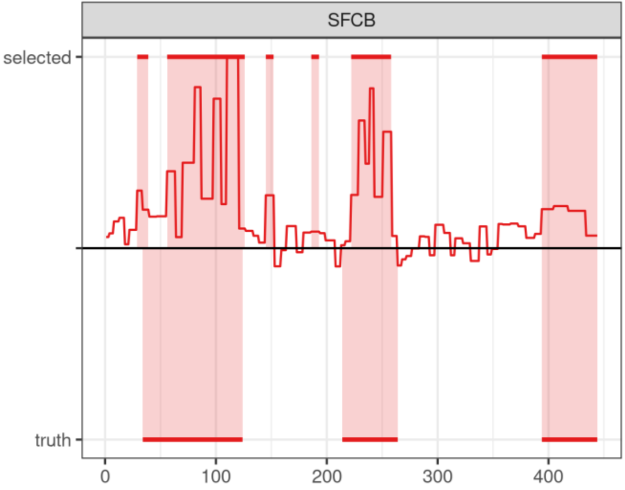


On simulated dataset

- X: 1,000 weather daily time series (WACSGen simulator [Flecher et al., 2010]) – $p = 444$
- Y:

$$y_i = \log\left(1 + |\langle x_i, \beta \rangle|\right) + \epsilon_i,$$

  - β: piecewise constant as "truth" on the left
  - $\epsilon_i \sim \mathcal{N}(0, 0.5)$

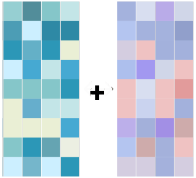


End of the story!

Questions?

Boruta

Back



RF + VI computation

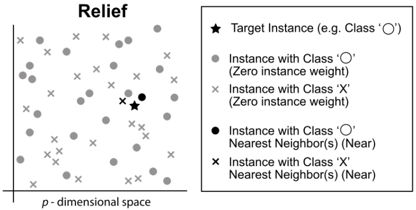
compare:

$\mathrm{VI}(j)$ $\max_j \mathrm{VI}(j)$

```
repeat
    augment dataset with randomized shadow copies

    train random forest and compute VI
    compute ZMOCK := max_{j: copies} VI(j)
    Decision: for j in initial variables
    if VI(j) > ZMOCK then
        j is selected
    else
        if VI(j) rejected by Student's test for ZMOCK then
            j is rejected
        end if
    end if
until all variables have been given a decision or Tmax iterations
have been performed
```

Relief



Iterative computation of weights:

1. pick an observation i at random
2. update weights of variable j:

$$w_j = w_j - (x_{ij} - x_{\text{nearest hit},j})^2 + (x_{ij} - x_{\text{nearest miss},j})^2$$

Back

Credits

- Evolution of the number of publications on time series has been obtained from `https://app.dimensions.ai`
- corn harvest image is "récolte du maïs à Épône (Yvelines)" by Spedona, from Wikimedia Commons
- DTW image is courtesy of Charlotte Pelletier
- Proximity forest split image is taken from [**Lucas et al., 2019**]
- Design of experiment impage for vaccine trial is taken from [**Capitaine et al., 2020**]
- Fréchet tree, related time series and Fréchet forest images are courtesy of Robin Genuer (and adapted to my needs)
- BOSS images are taken from [**Schäfer, 2015**]
- tree recoding image is taken from [**Baydogan and Runger, 2015**]
- black truffle basket image is "A basket of Summer black truffles from Mercato Gourmet by Giando" by Peachyeung316, from Wikimedia Commons
- Relief method image is "Illustration of Relief neighbor selection for scoring." by Docurbs from Wikimedia Commons

The rest is my own work.

References

(unofficial) Beamer template made with the help of Thomas Schiex, Matthias Zytnicki and Andreea Dreau:
https://forgemia.inra.fr/nathalie.villa-vialaneix/bainrae

Ambroise, C., Dehman, A., Neuvial, P., Rigaill, G., and Vialaneix, N. (2019).
Adjacency-constrained hierarchical clustering of a band similarity matrix with application to genomics.
Algorithms for Molecular Biology, 14:22.

Baragatti, M., Grollemund, P., Montpied, P., Dupouey, J.-L., Gravier, J., Murat, C., and Le Tacon, F. (2019).
Influence of annual climatic variations, climate changes, and sociological factors on the production of the Périgord black truffle (tuber melanosporum vittad.) from 1903-1904 to 1988-1989 in the vaucluse (france).
Mycorrhiza, 29:113–125.

Baydogan, M. G. and Runger, G. (2015).
Learning a symbolic representation for multivariate time series classification.
Data Mining and Knowledge Discovery, 29:400–422.

Capitaine, L., Bigot, J., Thiébaut, R., and Genuer, R. (2020).
Fréchet random forests for metric space valued regression with non Euclidean predictors.
Preprint arXiv:1906.01741v2.

Chavent, M., Liquet, B., Kuentz-Simonet, V., and Saracco, J. (2012).
ClustOfVar: an R package for the clustering of variables.
Journal of Statistical Software, 50(13):1–16.

Degenhardt, F., Seifert, S., and Szymczak, S. (2019).

Evaluation of variable selection methods for random forests and omics data sets.
Briefings in Bioinformatics, 20(2):492–503.

Deng, H., Runger, G., Tuv, E., and Martyanov, V. (2013).
A time series forest for classification and feature extraction.
Information Science, 239:142–153.

Flecher, C., Naveau, P., Allard, D., and Brisson, N. (2010).
A stochastic daily weather generator for skewed data.
Water Resources Reasearch, 46(7):W07519.

Fréchet, M. M. (1906).
Sur quelques points du calcul fonctionnel.
Rendiconti del Circolo Matematico di Palermo, 22:1–72.

Genolini, C., Ecochard, R., Benghezal, M., Driss, T., Andrieu, S., and Subtil, F. (2016).
kmlshape: an efficient method to cluster longitudinal data (time-series) according to their shapes.
PLoS ONE, 11(6):e0150738.

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2015).
VSURF: an R package for variable selection using random forests.
The R Journal, 7(2):19–33.

Keogh, E. J. and Pazzani, M. J. (2001).
Derivative dynamic time warping.
In Kumar, V. and Grossman, R., editors, *Proceedings of 2001 SIAM International Conference on Data Mining (SDM 2001)*, pages 1–11, Chicago, IL, USA. SIAM.

Kira, K. and Rendell, L. A. (1992).

A practical approach to feature selection.
In Sleeman, D. and Edwards, P., editors, *Proceedings of the International Conference on Machine Learning (ICML 1992)*, pages 249–256. Morgan Kaufmann Publishers, San Francisco, CA, USA.

Kursa, M. and Rudnicki, W. (2010).
Feature selection with the Boruta package.
Journal of Statistical Software, 36(11):1–13.

Lefort, G., Liaubet, L., Marty-Gasset, N., Canlet, C., Vialaneix, N., and Servien, R. (2021).
Joint automatic metabolite identification and quantification of a set of ^{1}H NMR spectra.
Analytical Chemistry, 93(5):2861–2870.

Lines, J., Taylor, S., and Bagnall, A. (2018).
Time series classification with HIVE-COTE: the hierarchical vote collective of transformation-based ensembles.
ACM Transactions on Knowledge Discovery from Data, 12(5):1–35.

Lucas, B., Shifaz, A., Pelletier, C., O'Neill, L., Zaidi, N., Goethals, B., Petitjean, F., and Webb, G. I. (2019).
Proximity forest: an effective and scalable distance based classifier for time series.
Data Mining and Knowledge Discovery, 33:607–635.

Massoni, S., Olteanu, M., and Villa-Vialaneix, N. (2013).
Which distance use when extracting typologies in sequence analysis? An application to school to work transitions.
In *International Work Conference on Artificial Neural Networks (IWANN 2013)*, Puerto de la Cruz, Tenerife.

Middlehurst, M., Large, J., and Bagnall, A. (2020).
The canonical interval forest (CIF) classifier for time series classification.
In Wu, X., Jermaine, C., Hu, X., Kotevskia, O., Lu, S., Xu, W., Aluru, S., Zhai, C., Al-Masri, E., Chen, Z., and Saltz, J., editors, *Proceedings of IEEE International Conference on Big Data*, Atlanta, GA, USA. IEEE.

Neethirajan, S. (2020).
Transforming the adaptation physiology of farm animals through sensors.
Animals, 10(9):1512.

Peterson, A. and Müller, H.-G. (2019).
Fréchet regression for random objects with euclidean predictors.
The Annals of Statistics, 47(2):691–719.

Picheny, V., Servien, R., and Villa-Vialaneix, N. (2019).
Interpretable sparse sliced inverse regression for functional data.
Statistics and Computing, 29(2):255–267.

Poterie, A., Dupuy, J.-F., Monbet, V., and Rouvière, L. (2019).
Classification tree algorithm for grouped variables.
Computational Statistics, 34:1613–1648.

Ramsay, J. and Silverman, B. (2002).
Applied Functional Data Analysis.
Springer Verlag.

Ramsay, J. and Silverman, B. (2005).
Functional Data Analysis.
New York, 2nd edition edition.

Randriamihamison, N., Vialaneix, N., and Neuvial, P. (2021).
Applicability and interpretability of Ward's hierarchical agglomerative clustering with or without contiguity constraints.
Journal of Classification, 38:363–389.

Robnik-Šikonja, M. and Kononenko, I. (1997).

An adaptation of Relief for attribute estimation in regression.
In Fisher, D. H., editor, *Proceedings of the International Conference on Machine Learning (ICML 1997)*, pages 296–304. Morgan Kaufmann Publishers, San Francisco, CA, USA.

Sakoe, H. and Chiba, S. (1978).
Dynamic programming algorithm optimization for spoken word recognition.
IEEE Transactions on Acoustics, Speech, and Signal Processing, 26(1).

Schäfer, P. (2015).
The BOSS is concerned with time series classification in the presence of noise.
Data Mining and Knowledge Discovery, 29(6):1505–1530.

Servien, R. and Vialaneix, N. (2023).
A random forest approach for interval selection in functional regression.
Preprint submitted for publication.

Shifaz, A., Pelletier, C., Petitjean, F., and Webb, G. J. (2020).
Ts-chief: a scalable and accurate forest algorithm for time series classification.
Data Mining and Knowledge Discovery, 34:742–775.

Speiser, J. L., Miller, M. E., Tooze, J., and Ip, E. (2019).
A comparison of random forest variable selection methods for classification prediction modeling.
Expert Systems with Applications, 134:93–101.

Studer, M. and Ritschard, G. (2016).
What matters in differences between life trajectories: a comparative review of sequence dissimilarity measures.
Journal of the Royal Statistical Society, Series A (Statistics in Society), 179(2):481–511.