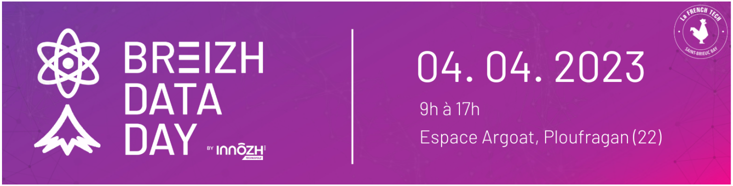
Modelling a collection of curves using functional data analysis.

Hassan Maissoro & Sunny Wang

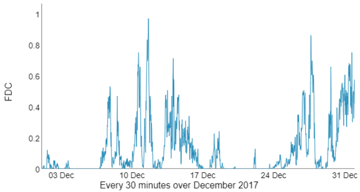


Outline

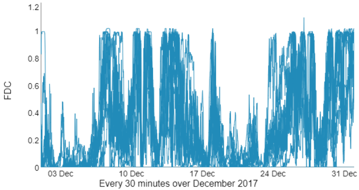
Business context for Datastorm

- Exploring the potential of FDA.
  - The paradigm is different from that of time series ;
  - The unit of observation is a curve or a vector of curves ;

Load curve of one wind farm



Load curves of some wind farms



- Applications in **medicine**, **meteorology**, **energy**, **finance**, etc.
- **The objective :** be able to deploy an FDA solution even when data is irregular.

Statistical issues

Example of electricity production of wind farms

The dimension increases exponentially.

- For each park, electricity production is recorded every **30 minutes** for more than 3 years.
  - That is **17,520 observations** per wind farm in 1 year, **35,040 observations** in 2 years...

Difficult to stay within the multivariate framework

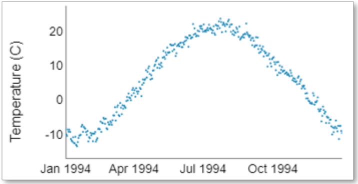
- There are approximately **1,500 wind farms**.
  - **1,500 wind farms** × **17 520 obs. times** over 1 year ;
  - It is difficult to take time dependency into account.

> Need for a new analytical framework : Functional Data Analysis (FDA).

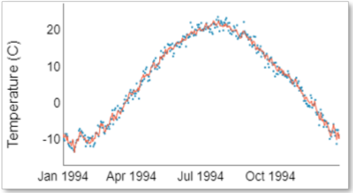
FDA overview

- Transforming discrete data to curves : non-parametric smoothing.

MONTRÉAL TEMPERATURE AVERAGED OVER 1960 TO 1994 + GAUSSIAN ERROR

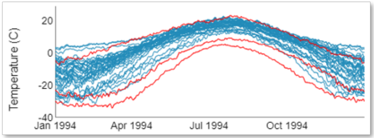


MONTRÉAL TEMPERATURE AVERAGED OVER 1960 TO 1994 + GAUSSIAN ERROR AND TRUE DATA



- Perform analyses :
  - Estimation of mean ;
  - Estimation of covariance ;
  - Anomaly detection ;
  - Robust prediction models ;
  - Etc.

CANADIAN TEMPERATURE AVERAGED OVER 1960 TO 1994 AT 35 LOCATIONS

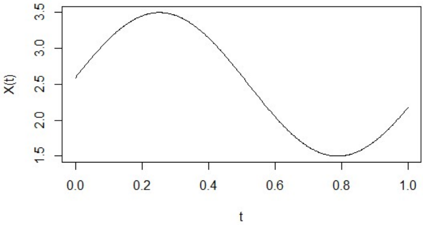


Functional data analysis framework

Functional data analysis framework (1/3)

- The random variables take values in a function space, usually Hilbertian, of infinite dimension.
- We suppose N i.i.d. observations $\{\mathrm{X}_i(t) : t \in [0,1], 1 \leq i \leq N\}$ of a random function $\mathrm{X} = \{\mathrm{X}(t) : t \in [0,1]\}$.
- A framework analysis different from time series.

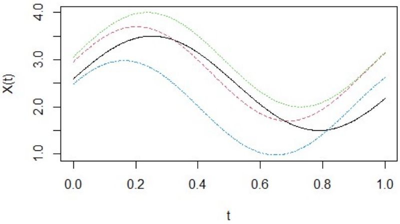
Figure – Time series, single trajectory $\{\mathrm{X}_1(t) : t \in [0,1]\}$.



Functional data analysis framework (1/3)

- The random variables take values in a function space, usually Hilbertian, of infinite dimension.
- We suppose N i.i.d. observations $\{\mathrm{X}_i(t) : t \in [0,1], 1 \leq i \leq N\}$ of a random function $\mathrm{X} = \{\mathrm{X}(t) : t \in [0,1]\}$.
- A framework analysis different from time series.

Figure – Functional data, trajectory collection $\{\mathrm{X}_i(t) : t \in [0,1], 1 \leq i \leq 4\}$.

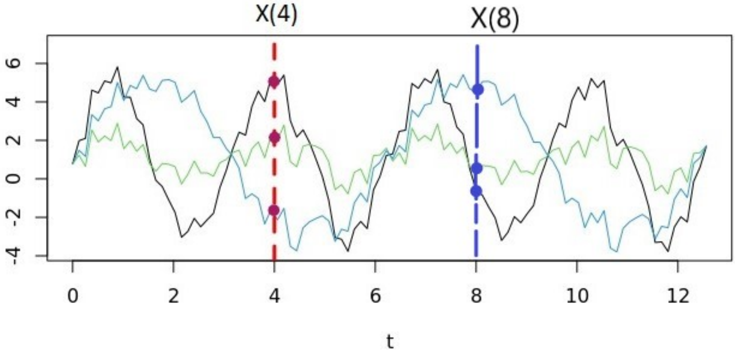


Functional data analysis framework (2/3)

A random function $\mathrm{X} = \{\mathrm{X}(t) : t \in [0,1]\} \in \mathrm{L}^2$ is mainly characterized by its mean function and its covariance function.

$$\mu(t) = \mathbb{E}\left[\mathrm{X}(t)\right] \text{ et } c(t,s) = \mathbb{E}\left[\Big(\mathrm{X}(t) - \mu(t)\Big)\Big(\mathrm{X}(s) - \mu(s)\Big)\right], \quad t,s \in [0,1].$$

Figure – Functional data, trajectory collection $\{\mathrm{X}_i(t) : t \in [0,13], 1 \leq i \leq 3\}$.



Functional data analysis framework (3/3)

- Moreover, we have the Karhunen-Loève expansion

$$\mathrm{X}_i(t) \;=\; \mu(t) \;+\; \sum_{j=1}^{\infty} \underbrace{\langle \mathrm{X}_i - \mu, v_j \rangle}_{\xi_{ij}=\text{score}} v_j(t), \quad t \in [0,1]$$

  with $\{v_1, v_2, \ldots\, v_j, \ldots\}$ an orthonormal basis of $\mathbb{L}^2$.
- If we know the scores, we can approximate $\mathrm{X}_i(t) - \mu(t)$ by

$$\left(\xi_{i1}, \xi_{i2}, \ldots, \xi_{ip}\right)^\top \in \mathbb{R}^p.$$

- For all t, a natural estimator of $\mu(t)$ is

$$\hat{\mu}(t) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{X}_i(t).$$

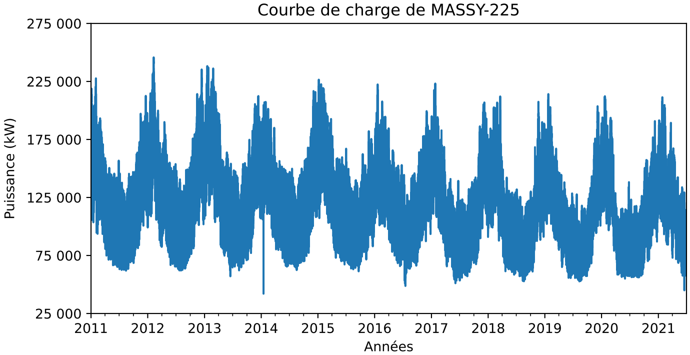
- We can estimate the scores ξ_{ij} using functional principal component analysis (FPCA).

Functional time series

Functional time series (1/3)

Example of a connection point for the extraction and injection of electricity

- A set of time-dependent curves.

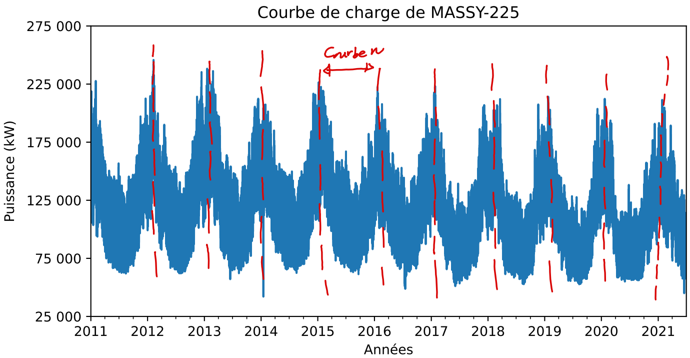


- This series can be decomposed into an annual profile, a weekly profile, a weather profile and a hazard profile.

Functional time series (1/3)

Example of a connection point for the extraction and injection of electricity

- A set of time-dependent curves.

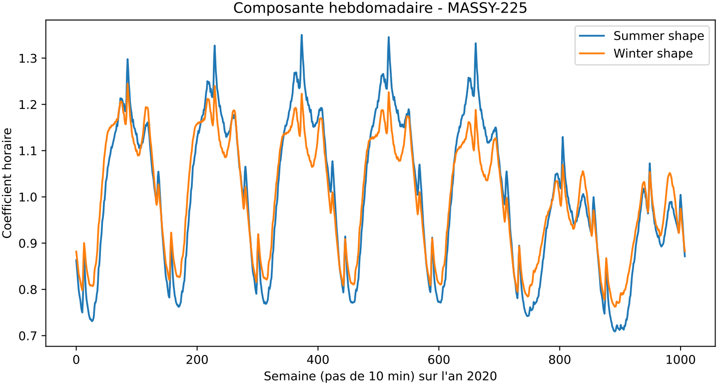


- This series can be decomposed into an annual profile, a weekly profile, a weather profile and a hazard profile.

Functional time series (2/3)

Example of a connection point for the extraction and injection of electricity

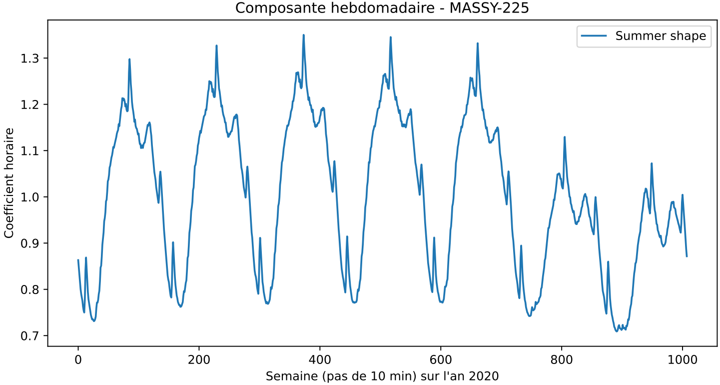
- A set of time-dependent curves.
- Focus on the weekly profiles : irregular curves.



Functional time series (3/3)

Irregular trajectories must be reconstructed...

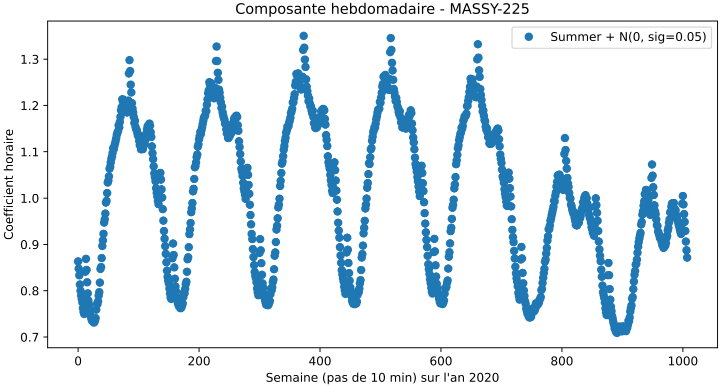
- Focus on the weekly summer profile : irregular curve.
- We observe the trajectory every 10 mins + measurement errors.
- Trajectories need to be reconstructed : an essential step in FDA..



Functional time series (3/3)

Irregular trajectories must be reconstructed...

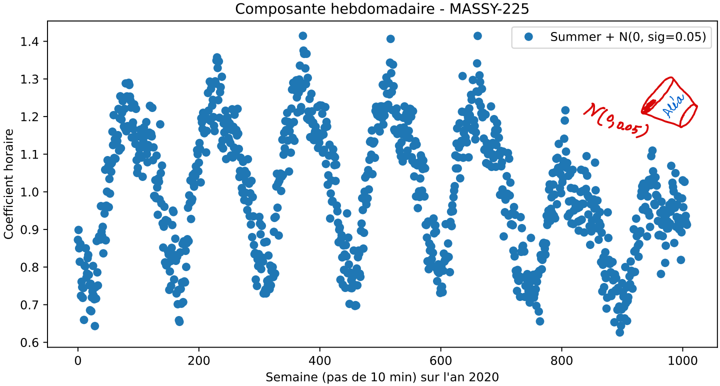
- Focus on the weekly summer profile : irregular curve.
- We observe the trajectory every 10 mins + measurement errors.
- Trajectories need to be reconstructed : an essential step in FDA..



Functional time series (3/3)

Irregular trajectories must be reconstructed...

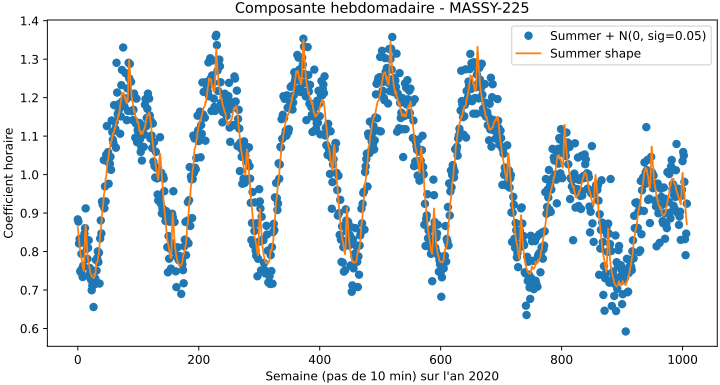
- Focus on the weekly summer profile : irregular curve.
- We observe the trajectory every 10 mins + measurement errors.
- Trajectories need to be reconstructed : an essential step in FDA..



Functional time series (3/3)

Irregular trajectories must be reconstructed...

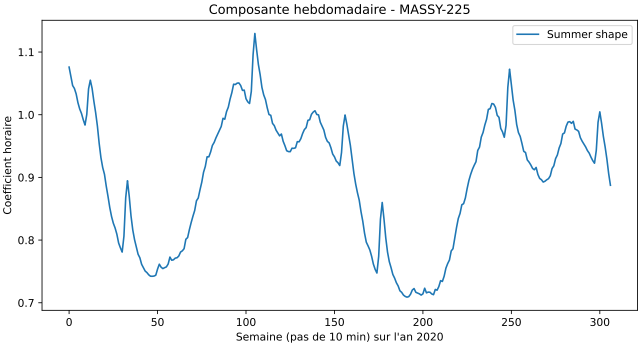
- Focus on the weekly summer profile : irregular curve.
- We observe the trajectory every 10 mins + measurement errors.
- Trajectories need to be reconstructed : an essential step in FDA..



Estimation of local regularity

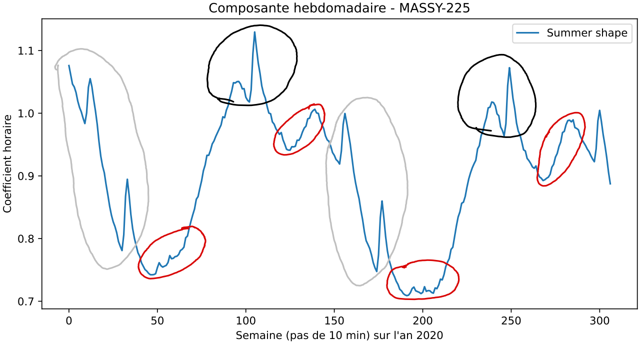
Motivation (1/2)

- Focus on the last few weeks of the summer 2020 weekly profile.
- There are several areas with different variations.
- The regularity of the trajectory varies locally.
- An optimal reconstruction should consider the local regularity.



Motivation (1/2)

- Focus on the last few weeks of the summer 2020 weekly profile.
- There are several areas with different variations.
- The regularity of the trajectory varies locally.
- An optimal reconstruction should consider the local regularity.



Motivation (2/2)

We aim to estimate the local regularity parameters of the trajectories for FTS (resp. i.i.d.) in the context of weak dependency.

Using dependent curves measured with noise at random discrete points, our goal is to perform adaptive estimation of :

- mean and covariance functions,
- auto-covariance function,
- depth functions,
- functional principal components, *etc*.

The concept of local regularity, considered by GOLOVKINE ET AL., (2022) for i.i.d. functional data, allows such constructions.

Definition of the local regularity (1/2)

The process X admits a *local regularity* at $t \in I$, with local exponent $H_t \in (0, 1)$ and Hölder constant $L_t > 0$, if

$$\mathbb{E}\left[(X(u) - X(v))^2\right] \approx L_t^2 |u - v|^{2H_t},$$

for all u, v satisfying $t - \Delta/2 \leq u \leq t \leq v \leq t + \Delta/2$ for some $\Delta > 0$.

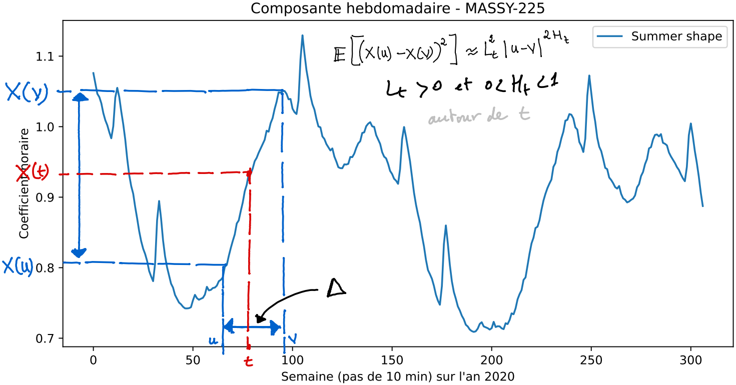
Data observation framework for the estimation. For $n = 1, \ldots N$, X_n is measured with error at discrete, randomly sampled points :

$$Y_{n,k} = X_n(T_{n,k}) + \varepsilon_{n,k}, \quad 1 \leq k \leq M_n,$$

- $M_1, \ldots, M_N \overset{i.i.d.}{\sim} M$ with expectation μ,
- the observation times $T_{n,k} \sim T$ are independent,
- $\varepsilon_{n,k} \sim \epsilon$ are independent centered errors,
- $\{X_n\}$, M, ϵ, and T are mutually independent.

Definition of the local regularity (2/2)

- How to choose Δ ? How many curves and how many points per curve are needed to make it work ?
- The answer is given by concentration bounds under the assumption of $\mathbb{L}^{\mathbf{4}}_{\mathcal{C}}$ − **m-approximable** for FTS.



Estimation of local regularity parameters

We use some nonparametric estimates $\widetilde{X}_n$ to recover the X_n's.

For any u, v close to t, let

$$\widehat{\theta}(u,v) = \frac{1}{N}\sum_{n=1}^{N}\left\{\widetilde{X}_n(v) - \widetilde{X}_n(u)\right\}^2.$$

Our estimators of H_t and L_t^2 are defined as empirical counterparts of their respective definition.

Let $t_1 = t - \Delta/2$, $t_3 = t + \Delta/2$. The estimators of H_t and L_t^2 are

$$\widehat{H}_t = \frac{\log(\widehat{\theta}(t_1,t_3)) - \log(\widehat{\theta}(t_1,t))}{2\log(2)} \qquad \text{and} \qquad \widehat{L}_t^2 = \frac{\widehat{\theta}\,(t_1,t_3)}{\Delta^{2\widehat{H}_t}}.$$

Application (1/2)

Estimation of local regularity parameters

We simulate a FAR(1) where $\{\varepsilon_n\}$ are i.i.d. 'tied-down' *multifractional Brownian motion* (see Stoev and Taqqu (2006)) paths with :

- a logistic H_t function and $L_t^2 = 4$,
- and a kernel $\beta(s, t) = \alpha st$, with $\alpha = 9/4$.

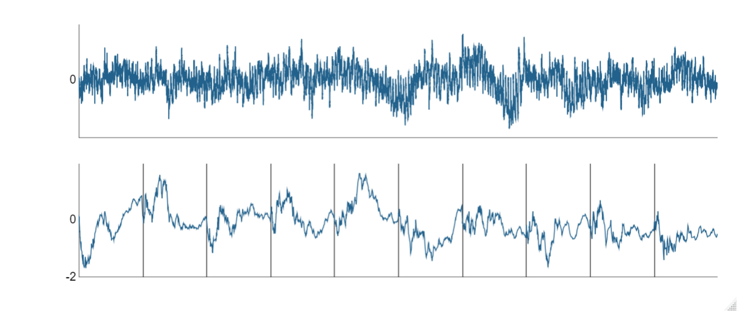


Figure – Time series of $N = 250$ observations of a simulated FAR(1) without error. The last ten functions are shown in the bottom graph.

Application (2/2)

Estimation of local regularity parameters

Estimation of H_t and L_t^2 at $t = 1/2$ using the previous FAR(1) and taking $\epsilon \sim \mathcal{N}(0, 0.04)$.

- Obtained reasonably good results :

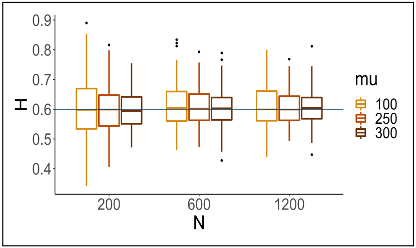


Figure – Estimates of $\widehat{H}_t$. The line is the true $H_t = 0.6$.

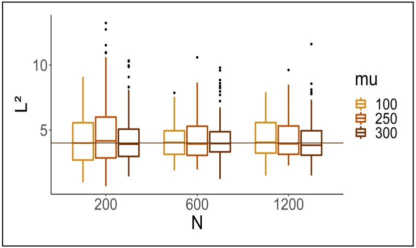


Figure – Estimates of $\widehat{L}_t^2$. The line is the true $L_t^2 = 4$.

Functional Principal Components Analysis (FPCA)

Motivation (1/2)

- In practice, all supposedly functional data are observed as discrete points along a curve.
- Observations can be densely observed on an equally spaced grid (e.g sensors data), or more sparsely observed at random time points with large spacings (e.g longitudinal data in clinical trials).
- In any case, curves need to be reconstructed from these discrete points - this process is called *smoothing*.
- Smoothing also plays the dual role of noise removal.

Motivation (2/2)

- To avoid placing parametric assumptions, smoothing is usually performed using non-parametric regression smoothers.

- Common choices include local polynomial estimators, smoothing splines etc.

- In any case, they will require the selection of a smoothing parameter.

- **Key Question : How should the smoothing parameter be chosen ?**

- Important because it affects the inference later !

Motivation (3/3)

- Recent literature (Golovkine et al. (2021)) have shown that the optimal choice of smoothing parameter largely depends on the purpose .

- For example, the degree of smoothing required is generally different for optimally computing the mean vs the covariance function (i.e. “Purpose-driven smoothing").

- Especially the case for sparsely sampled curves observed at random time points.

A detour (sort of...) to FPCA

- FPCA is a fundamental tool in fda, because it allows one to perform the finite dimensional analysis of a problem that is intrinsically infinite dimensional.
- After smoothing, FPCA is usually the next practical step, since further analyses such as regression is usually performed using the principal components.
- Construction of the principal components require smoothed curves $\Longrightarrow$ smoothing parameter again need to be selected !

State of affairs - FPCA

- fda literature is largely silent on the principled choice of smoothing parameter for the purposes of FPCA.
- Usually involves some combination of ad-hoc judgement and automatic, data-driven methods such as cross-validation.
- Unfortunately, cross-validation is very computationally expensive, especially in the context of FPCA if one wants to use the information present in all the curves.
- Good rule for selecting smoothing parameter should use all signals carried by curves, be computationally efficient, and automatically adjust to the sampling scheme (dense vs sparse).
- Difficult problem, but we propose such a rule for smoothing parameter selection that satisfies all these properties ! !

Setup

- Same data setup as regularity estimation : observations are pairs $(Y_{n,m}, T_{n,m}) \in \mathbb{R} \times \mathcal{T}$ such that

$$Y_{n,m} = X_n(T_{n,m}) + \varepsilon_{n,m}, \tag{1}$$

  with noise $\varepsilon_{n,m}$ possibly heteroscedastic.
- Focus on local polynomial estimators, more specifically the Nadaraya-Watson estimator

$$\widehat{X}_n(t;h) = \sum_{m=1}^{M_n} W_{n,m}(t;h) Y_{n,m}, \tag{2}$$

  where the weights $W_{n,m}$ are such that

$$W_{n,m}(t;h) = \left(\sum_{m=1}^{M_n} K\left(\frac{T_{n,m} - t}{h} \right) \right)^{-1} K\left(\frac{T_{n,m} - t}{h} \right). \tag{3}$$

- Smoothing parameter is the bandwidth h.

Key Idea

- Ability to propose a good bandwidth rule is due to the derivation of explicit quadratic risk bounds.
- Risk bounds are adapted to each eigen-element : so we have a different risk function $\mathcal{R}_N(\lambda_j; h)$ and $\mathcal{R}_N(\psi_j; h)$ for each index j.
- Our bandwidth rule minimises these risk bounds :

$$h^*(\lambda_j) = \arg \min_{h \in \mathcal{H}_N} \mathcal{R}_N(\lambda_j; h), \tag{4}$$

and

$$h^*(\psi_j) = \arg \min_{h \in \mathcal{H}_N} \mathcal{R}_N(\psi_j; h). \tag{5}$$

The FPCA algorithm

1. Estimate key parameters to compute risk bounds, such as H and L.
2. Numerical computation of risk bounds $\mathcal{R}_N(\lambda_j; h)$ and $\mathcal{R}_N(\psi_j; h)$. Obtain h^* that minimises them.
3. Constructing smoothed curves. Plug in h^* into (2).
4. Constructing covariance estimates $\widehat{\Gamma}(s, t; h^*(\lambda_j))$ and $\widehat{\Gamma}(s, t; h^*(\psi_j))$ for each eigen-element, using

$$\widehat{\Gamma}(s,t;h)=\frac{\sum_{n=1}^{N} w_n(s,t;h)\left\{\widehat{X}_n(t;h)-\widehat{\mu}_N(t;h)\right\}\left\{\widehat{X}_n(s;h)-\widehat{\mu}_N(s;h)\right\}}{\mathcal{W}_N(s,t;h)}. \tag{6}$$

5. Perform eigen-analysis of $\widehat{\Gamma}$'s to obtain $\widehat{\lambda}_j$ and $\widehat{\psi}_j$.

Some comments

- Why can our bandwidth rule satisfy the desired properties ?
  1. Minimisation of risk bounds $\Longrightarrow$ optimality.
  2. Explicit nature of risk bounds $\Longrightarrow$ computational efficiency.
  3. Usage of regularity properties of curves $\Longrightarrow$ all signals are exploited !
- Theoretical guarantees are also proved.

Application setup (1/2)

- Goal : Clustering of 1133 wind farms based on their energy load produced
- Study the daily average production over a period of one year. Each wind farm is a curve, with 365 points along each curve, sampled on a common grid
- Clustering outcomes using adaptive vs non-adaptive methods for FPCA are compared
- Non-adaptive comparison : perform penalised smoothing using a Fourier basis, and compute the empirical principal components
- Focus on Hierarchical agglomerative clustering (HAC), a commonly used method

Application setup (2/2)

- After FPCA is performed, we compute the estimated *scores*

$$\widehat{\xi}_{n,\ell} = \int_{\mathcal{T}} \left\{ \widehat{X}_n(t) - \widehat{\mu}(t) \right\} \widehat{\psi}_\ell(t) dt. \tag{7}$$

- Build a distance matrix D based on the Euclidean distance, with each component $d(i,j) = \sum_{\ell=1}^{L} \left(\widehat{\xi}_{i,\ell} - \widehat{\xi}_{j,\ell} \right)^2$. Apply HAC to D.
- Measure of performance : *silhouette* value

$$s(i) = \begin{cases} \frac{b(i)-a(i)}{\max\{a(i),b(i)\}}, & \text{if } |C_I| > 1 \\ 0 & \text{if } |C_I| = 1, \end{cases} \tag{8}$$

where $a(i) = (|C_I| - 1)^{-1} \sum_{j \in C_i : i \neq j} d(i,j)$, and
$b(i) = \min_{J \neq I} (|C_j|)^{-1} \sum_{j \in C_J} d(i,j)$.

Results (1/2)

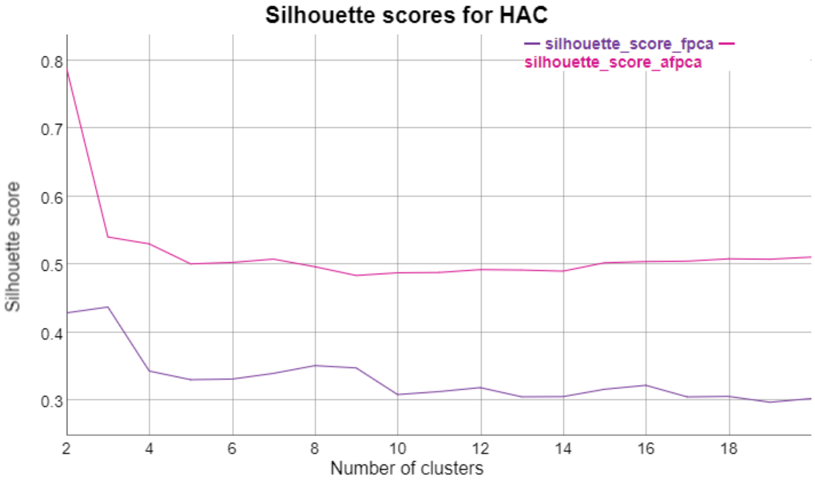


Figure – Silhouette scores for HAC

Results (2/2)

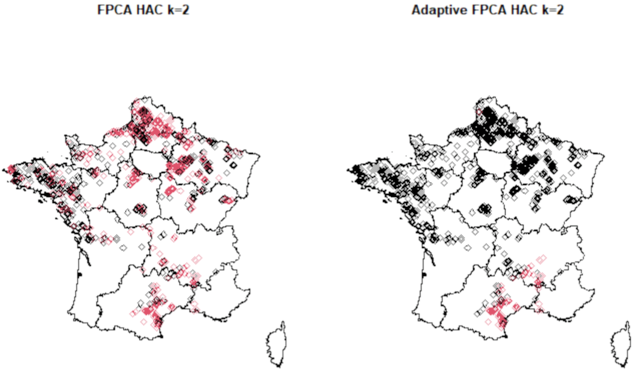


Figure – Clusters plotted on a map

Finishing thoughts

- A refined bandwidth rule for smoothing curves for the purposes of FPCA is proposed.
- Works well in practice : even if bandwidth rule was tailored for FPCA and not clustering !
- Should work even better if an adaptive bandwidth rule is derived specifically for clustering ! (future work perhaps ?)

Conclusion and perspectives

Conclusion and perspectives

- Global introduction to functional data analysis.
  - Different from time series analysis.
  - It is the analysis of a collection of dependent or independent curves.
- Estimation of local regularity.
  - Local regularity parameters are : exponent H_t and Hölder constant L_t^2.
  - The simulations show that $\widehat{H}_t$ and $\widehat{L_t^2}$ give satisfactory results.
- Functional principal components.
  - Optimal smoothing parameter used to reconstruct curves depends on the end goal of the practitioner.
  - For FPCA, a good bandwidth rule can be proposed, by building upon local regularity estimates.
- Perspectives :
  - adaptive estimators for functional time series analysis,
  - adaptive estimators for linear regression models,
  - adaptive estimators for anomaly detection, etc.

Thanks for your attention !