

# Introduction to functional data analysis

Universitat Autònoma de Barcelona



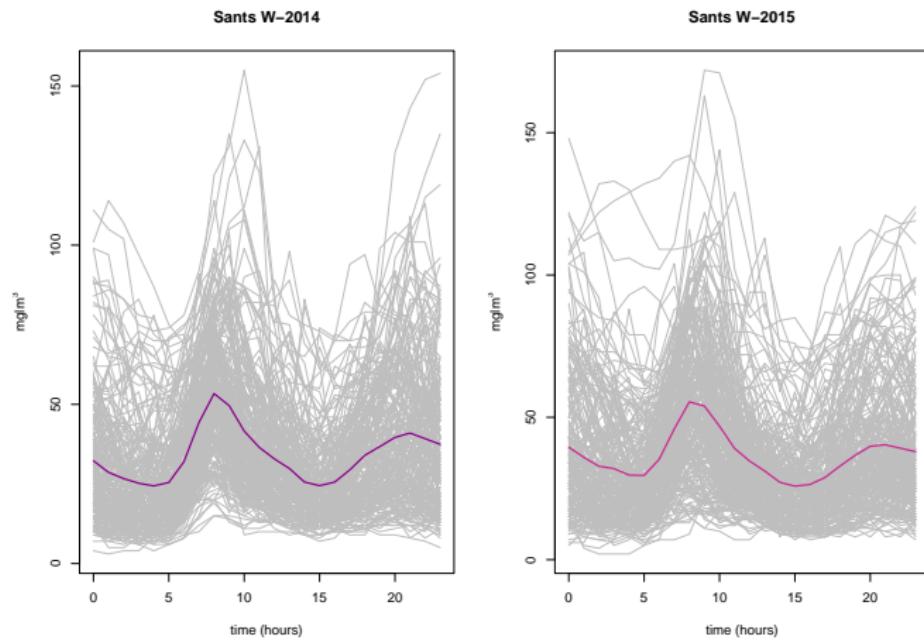
MBD 2018

- ① Examples of functional data
- ② What do we understand by Functional Data?
- ③ Descriptive Statistics
- ④ On Permutation tests
- ⑤ Two sample problem for functional data
- ⑥ Shilling's type test
- ⑦ Depths-based tests
- ⑧ Empirical comparison of powers
- ⑨ NO<sub>2</sub> Contamination in Barcelona

# Introduction to FDA

- Observing and saving complete functions as the result of random experiments is possible by the development of real-time measurement instruments and data storage resources.
- **Example:** For patients involved in a clinical trial, the blood pressure is monitored in continuous-time during 24 hours.
- Samples where a whole curve is observed at each sampling unit are referred to as Functional Data.
- Random functions are the statistical atoms in these cases.

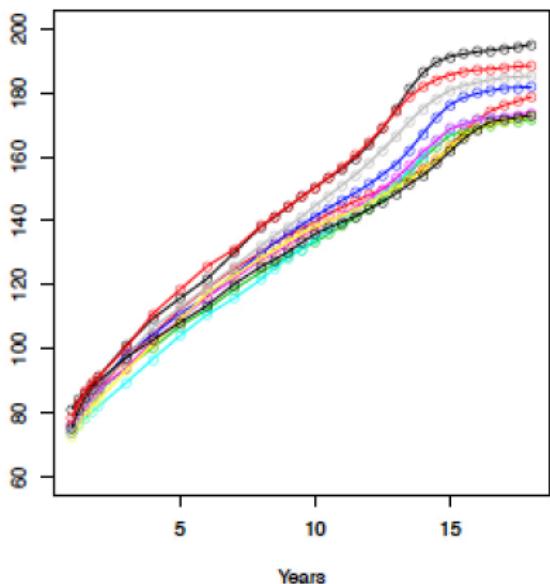
$\text{NO}_2$  levels on working days taken on the same station, in 2014 and 2015. Did (the law) of contamination- levels change?



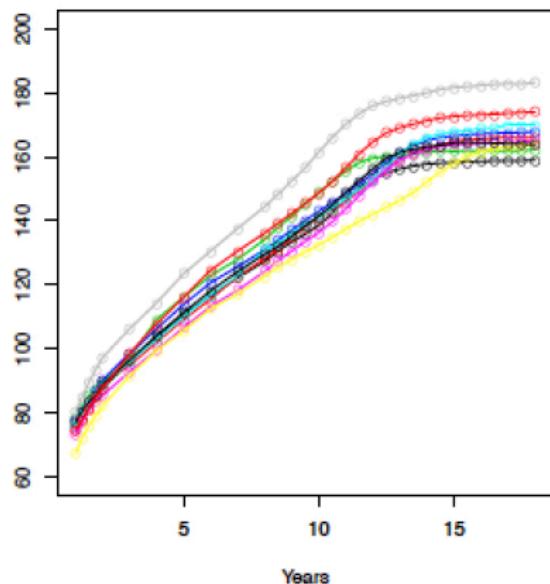
## Growth Curves

Heights of 39 boys and 54 girls measured at a set of 31 ages (from 1 to 18 years) in the Berkeley Growth Study. Only 10 are shown.

Height for males



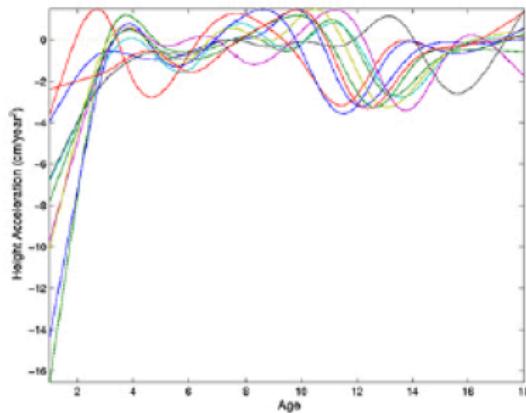
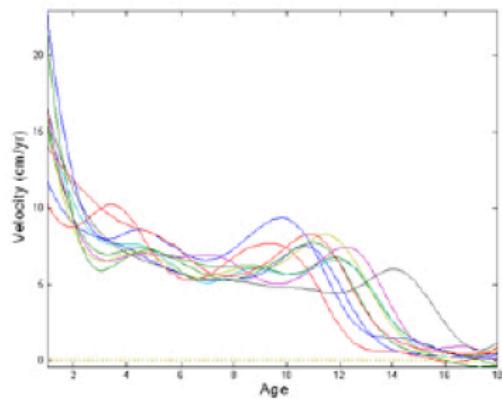
Height for females



```
> library(fda); data(growth); demo(growth)
```

# Growth Curves

The derivatives of functional data are new functional data, that can be as informative as the original functions (or even more!).

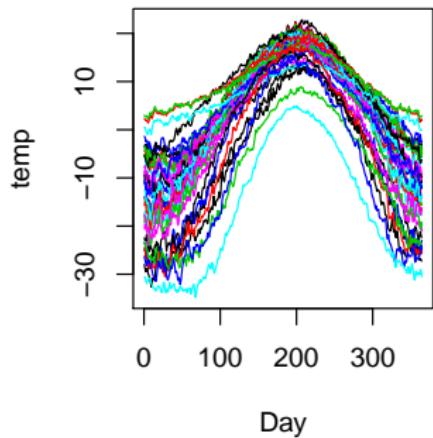


Height velocities and accelerations of 10 girls in the Berkeley Growth Study.

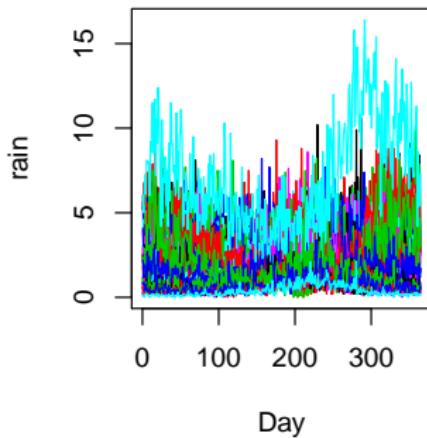
# Canadian Weather Data

This is a classical dataset, that consists of average (from 1960 to 1994) of daily temperature and precipitation records in 35 weather stations across Canada.

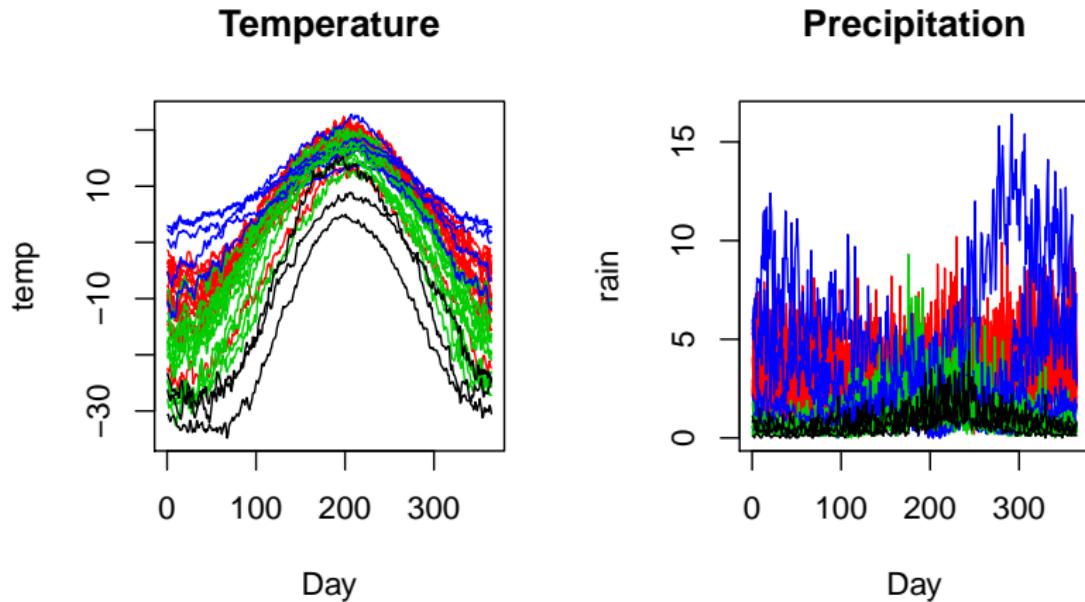
Temperature



Precipitation

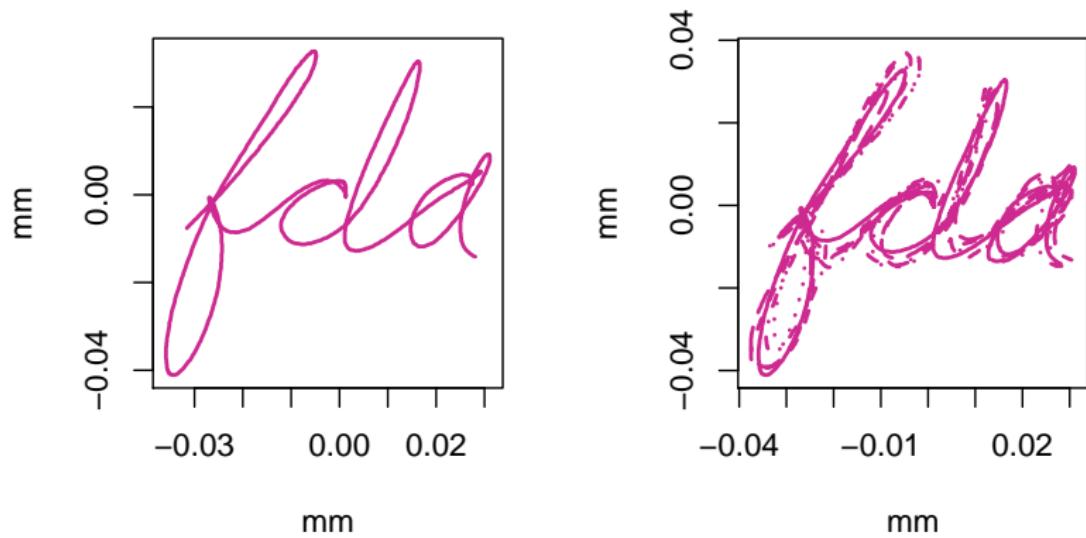


Different colors correspond to different geographical regions : Atlantic, Pacific, Central and North.



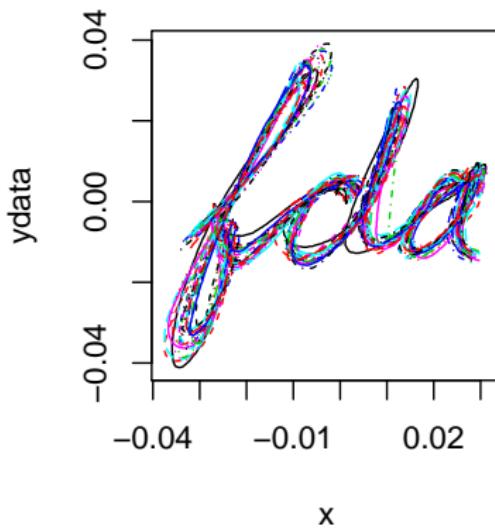
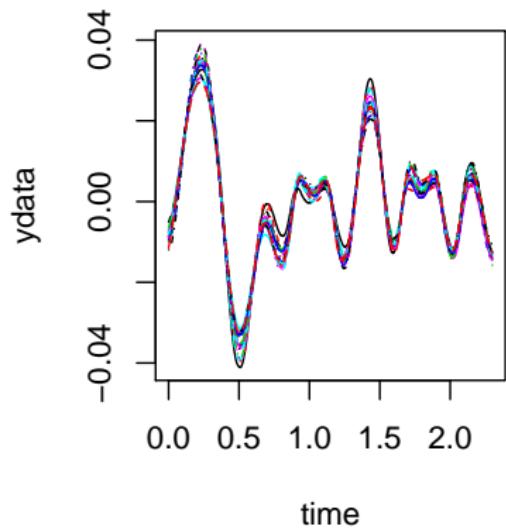
Interest is in variation and relationships between smooth underlying processes.

Multivariate functional data often arise from tracking movements of points through space.



This figure corresponds to the  $X - Y$  coordinates of 1 and 5 samples superimposed. The role of time is lost in this plots.

We can include the information of the time taken by Jim Ramsay to actually do the writing.



Each replication is represented by 1401 coordinate values. The scripts have been extensively pre-processed. They have been adjusted to a common length that corresponds to 2.3 seconds or 2300 milliseconds, and they have already been registered so that important features in each script are aligned.

## Some references

**Functional data Analysis** deals with the statistical description and modelling of random functions.

A broad outlook of FDA can be found in the books by

- Ramsay and Silverman (2002, 2005)
- Ramsay, Hooker and Graves (2009)
- Ferraty and Vieu (2006)
- Horvàth and Kokoszka (2012)
- Kokoszka and Reimherr (2017)
- R packages `fda` (Ramsay & Silverman) and `fda.usc` (Febrero Bande & de la Fuente)
- R task view <https://cran.r-project.org/web/views/FunctionalData.html>

## Functional data

We call functional data to a collection of observations of a random process that is registered as a function  $X_i : I \subset \mathbb{R} \rightarrow \mathbb{R}^n$ .

In practice, the concept of a continuum measurement doesn't exist. The process has to be monitored in a discrete grid  $t_1, t_2, \dots, t_L$ .

So at the end, one always has a vectorial observation  $(x(t_1), \dots, x(t_L))$ . There are several reasons to consider this as functional data:

- ① the continuous nature of the phenomenon at hand,
- ② the theoretical possibility of observing the phenomenon in a much finer grid,
- ③ the choice of a functional model to approximately represent it.

The key assumption is smoothness:

$$y_{i,j} = x_i(t_{i,j}) + \epsilon_{i,j}$$

with  $t$  in a continuum (usually time) and  $x_i(t)$  functions belonging to some “decent” functional space.

- In order to make inference on this type of data, we shall study later the space  $L^2(T, \lambda)$  of square integrable functions defined on a set  $T$  with respect to some measure  $\lambda$  (usually Lebesgue measure).
- But before that, we shall discuss several aspects of descriptive statistics for functional data.

# What are we interested in?

- Characteristics of data sets
  - mean
  - variation, covariation
  - a concept of “order”  $\Rightarrow$  “depth”
- Relationships of functional data to
  - covariate
  - responses
- Relationships between derivatives of functions.
- Timing of events in functions.

## What are the challenges?

- Estimation of functional data from noisy, discrete observations,
- Numerical representation of infinite-dimensional objects
- Representation of variation in infinite dimensions
- Description of statistical relationships between infinite dimensional objects.
- $n < p = \infty$ , and use of smoothness.
- Measures of variation in estimates.

## Statistical tools for FDA (so far...)

- Regression models: lm, glm, non-parametric regression, ANOVA...
- Multivariate Analysis: PCA, MDS, Clustering, Depth measures, ...
- Time Series, Spatial Statistics, ...
- Things that exploit the functional nature of the data
  - Principal differential analysis: a kind of PCA on the derivatives of the observed functions
  - Registration: a pre-process step that consists of a change of variables each observed function in order to make them as similar as possible.

# From Discrete to Functional Data

We need to represent data recorded at discrete times as continuous functions in order to

- Allow evaluation of records at any time point (specially if observation times are not the same across records).
- Evaluate rates of change (derivatives!)
- Reduce noise.
- Allow registration onto a common time-scale.

# From Discrete to Functional Data: representing non-parametric continuous time functions

## Basis-expansion methods

$$x(t) = \sum_{i=1}^K \phi_i(t)c_i$$

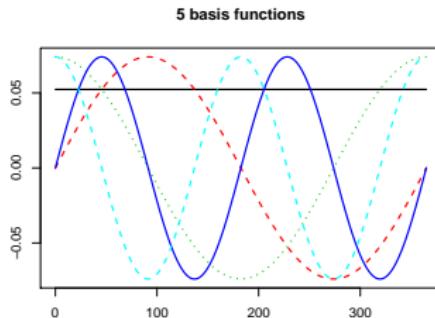
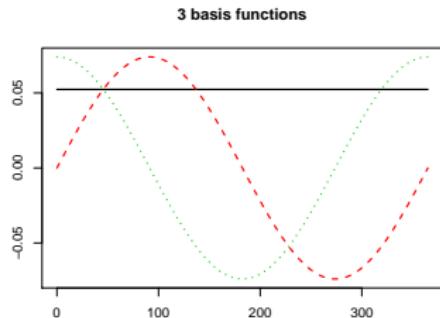
for pre-defined  $\phi_i$  and coefficients  $c_i$ . Several basis systems available:

- Fourier basis: ideal for periodic data, such as weather cycles, signal processing...
- B-splines
- wavelets, Monomials ( $1, x, x^2, \dots$ ), Power ( $t^{\lambda_i}$ ,  $\lambda_i$  not necessarily integer), Exponential ( $\exp(\lambda_i t)$ ).

# Fourier Basis

$$1, \sin(\omega t), \cos(\omega t), \sin(2\omega t), \cos(2\omega t), \dots, \sin(m\omega t), \cos(m\omega t), \dots$$

$\omega = 2\pi/P$  defines the period of oscillation of the 1st sine-cosine pair.

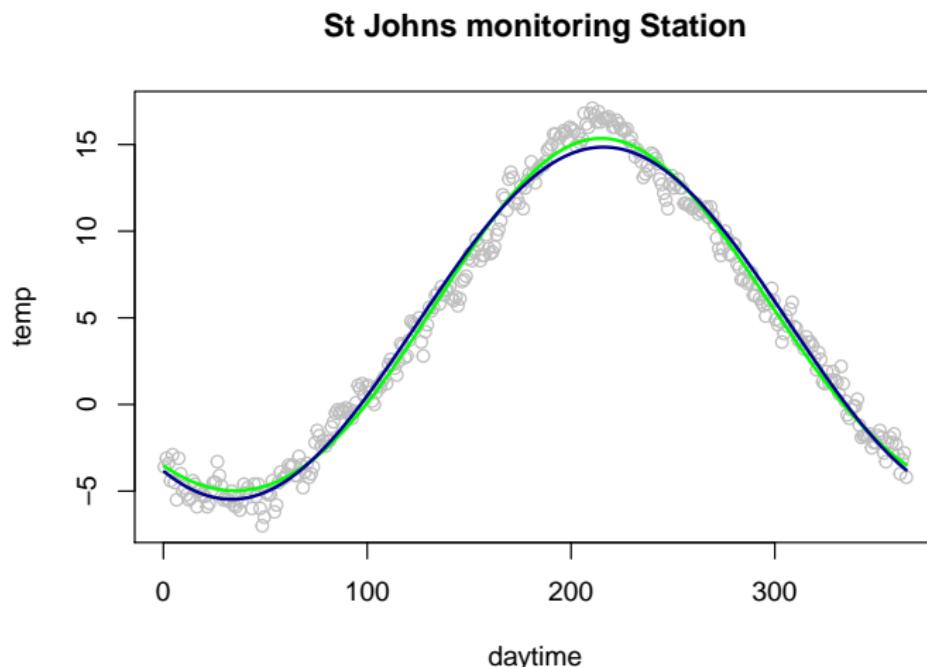


Natural for describing periodic data, such as annual weather cycles, but they always yield periodic functions, hence, it can be a problem in certain contexts.

## Example: Canadian Weather Data, St John's Station

Linear regression on temperature with the first 3 (blue) and 5 (green) basis functions.

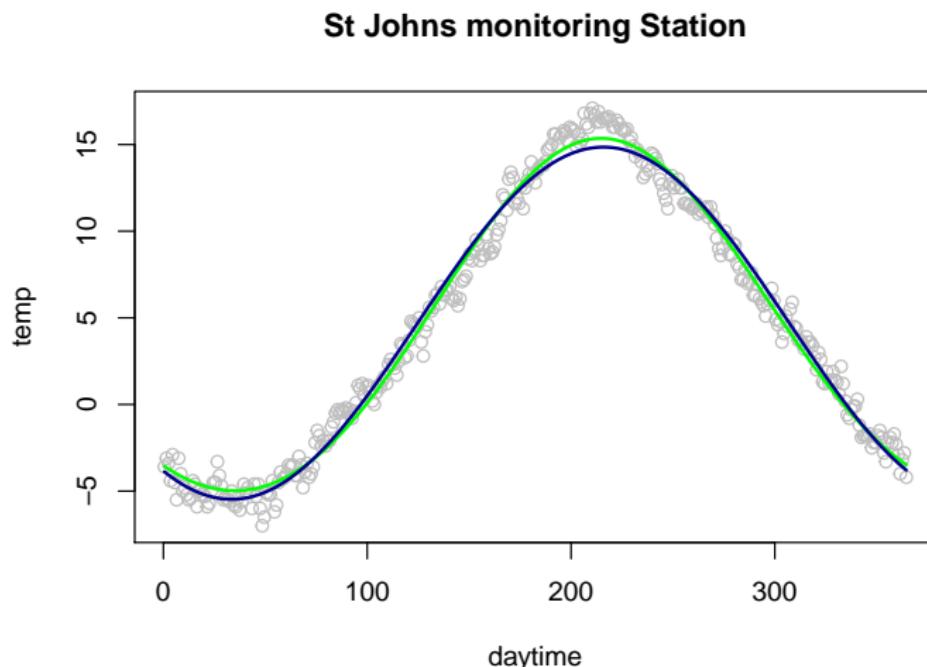
The basis functions are evaluated at the observation times



## Example: Canadian Weather Data, St John's Station

Linear regression on temperature with the first 3 (blue) and 5 (green) basis functions.

The basis functions are evaluated at the observation times



## Spline Basis (B-splines)

- Splines are polynomial segments joined end-to-end.
- Segments are constrained to be smooth at the joins.
- The points at which segments join are called *knots*.
- Systems are defined by
  - The order  $m$  (order=degree+1) of the polynomial.
  - The location of the knots.
- Number of basis functions:  
$$order + number\ of\ interior\ knots$$
- Sum of all B-splines in a basis is always 1; can fit any polynomial of order  $m$ .

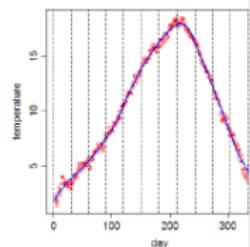
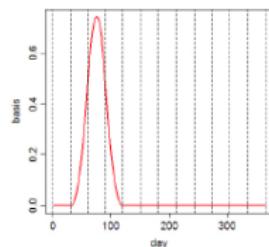
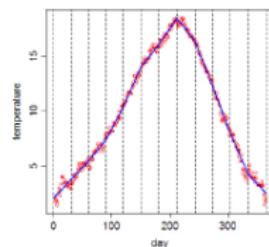
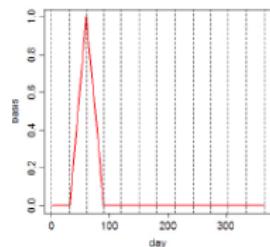
## Properties of B-Splines

- Order  $m$  splines: derivatives up to  $m - 2$  are continuous.
- Most popular choice is order 4, implying continuous second derivatives. Second derivatives have straight-line segments.
- B-spline basis functions are positive over at most  $m$  adjacent intervals  
⇒ fast computation for even thousands of basis functions.
- Support on  $m$  adjacent intervals: highly sparse design matrix.
- Flexibility comes from knots & derivatives from order.
- Frequently, fewer knots do reasonably well (approximation properties can be formalized).

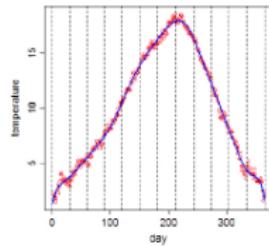
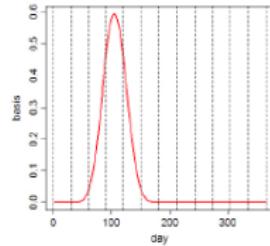
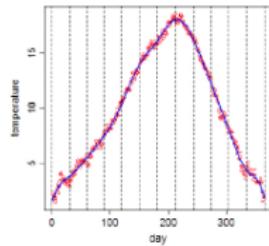
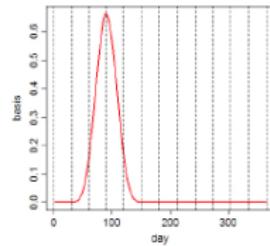
See de Boor, 2001, *A practical guide to splines*.

# Example: Canadian Weather Data, Vancouver

Splines of order 1 and 3, knots at months.



Splines of order 4 and 5, knots at months.



## B-splines: Choosing Knots and Order

- The order of the spline should be at least  $k + 2$  if you are interested in  $k$  derivatives..
- Knots are often equally spaced (a useful default).
- But there are two important rules:
  - Place more knots where you know there is strong curvature, and fewer where the function changes slowly.
  - Be sure there is at least one data point in every interval.
- Sum of all B-splines in a basis is always 1; can fit any polynomial of order  $m$ .

# Smoothing Penalties

Assume we have observations for a single curve

$$y_i := x(t_i) + \varepsilon \quad i = 1, \dots, N$$

and we want to estimate

$$x(t) \approx \sum_{j=1}^K c_j \phi_j(t).$$

Minimize the sum of squared errors:

$$SSE = \sum_{i=1}^N (y_i - x(t_i))^2 = \sum_{i=1}^n (y_i - c^T \Phi(t_i))^2$$

just as in linear regression.

If the  $N \times K$  matrix  $\Phi$  contains the values  $\phi_k(t_j)$  and  $\mathbf{y} = (y_1, \dots, y_n)$

$$SSE(c) = (\mathbf{y} - \Phi c)^T ((\mathbf{y} - \Phi c))$$

The error sum of squares is minimized by the OLS estimate

$$\hat{c} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}$$

and the estimate for  $x(t)$  is

$$\hat{x}(t) = \Phi(t)(\Phi^T \Phi)^{-1} \Phi^T \mathbf{y}.$$

- Least squares is optimal if residuals are iid  $N(0, \sigma^2)$ .
- If the variance is non-constant, weighted least squares can help.

$$WMSE(x) = \sum w_i(y_i - x(t_i))^2$$

If  $\mathbf{W} = \text{diag}(w_i)$ ,  $\hat{x}(t) = \Phi(t)\hat{c} = \Phi(t)(\Phi^T \mathbf{W} \Phi)^{-1} \Phi^T \mathbf{W} \mathbf{y}$

and at the observation points

$$\hat{\mathbf{y}} = \Phi(t)(\Phi^T \mathbf{W} \Phi)^{-1} \Phi^T \mathbf{W} \mathbf{y} = \mathbf{S} \mathbf{y}$$

$\mathbf{S}$  is called the smoothing matrix. So all we need to do is choosing the weights appropriately. This amounts to choosing the type of basis and the number of basis functions.

Trade off:

- Too many basis functions over-fit the data and reflect errors of measurement: small bias, large variance.
- Too few basis functions fail to capture interesting features of the curves: small variance, large bias.
- As often, cross-validation is a good idea.

## An alternative: Kernel methods

One ‘bump’ function  $K_i(t)$  at each data point calculate a weighted average

$$x(t) = \frac{\sum y_i K_i(t/h)}{\sum K_i(t/h)}$$

and use the band-with  $h$  to regulate the smoothness. Hence, we need a fair definition of “smooth”: what we really want is to eliminate “wiggles” in the data. Define the differential operator.

$$Dx(t) = \frac{d}{dt}x(t)$$

Higher order derivatives are defined in terms of powers:

$$D^2x(t) = \frac{d^2}{dt^2}x(t), \dots, D^kx(t) = \frac{d^k}{dt^k}x(t)$$

$Dx(t)$  is the slope of  $x(t)$ ,  $D^2x(t)$  is its curvature. And we measure the size of the curvature by

$$J_2 = \int (D^2x(t))^2 dt$$

# The smoothing Spline Theorem

Consider the penalized squared error<sup>1</sup>  $PENSSE_\lambda x(t)$  defined by

$$\sum (y_i - x(t_i))^2 + \lambda \int (D^2 x(t))^2 dt = (\mathbf{y} - x(\mathbf{t}))^T (\mathbf{y} - x(\mathbf{t})) + \lambda J_2$$

- $\lambda$  is a smoothing parameter measuring compromise between fit and smoothness.
- As  $\lambda$  increases, roughness is increasingly penalized and  $x(t)$  will become linear.
- As  $\lambda$  decreases, the penalty is reduced and allows  $x(t)$  to fit the data better.
- A remarkable theorem tells us that the function  $x(t)$  that minimizes  $PENSSE_\lambda(x)$  is a spline function of order 4 (piecewise cubic) with a knot at each sample point  $t_j$

---

<sup>1</sup> see Chapter 5 of Ramsay, Hooker, Graves(2009), Functional Data Analysis with R and Matlab, which is a sort of manual for package `fda`

## Calculating the penalized fit

- The theorem tells us that  $x(t)$  is of the form  $x(t) = \Phi(t)^T c$  where  $\phi(t)$  is a vector of B-spline basis functions.
- The number of basis functions is  $(n - 2) + 4 = n + 2$  where  $n$  is the number of sampling points.

Using that  $x(t) = \Phi(t)c$ , we have:

$$\int (D^2 x(t))^2 dt = \int (c^T (D^2 \phi(t)) (D^2 b m \phi(t))^T c dt = c^T R_2 C$$

where  $[R_2]_{i,j} = \int D^2 \phi_i(t) D^2 \phi_j(t) dt$  is the penalty matrix (2 can be changed by  $m$ ). The penalized least squares estimate of  $c$  is

$$\hat{c} = [\Phi^T \Phi + \lambda R_2]^{-1} \Phi^T y$$

and the linear smoother is

$$\hat{y} = \phi [\Phi^T \Phi + \lambda R_2]^{-1} \Phi^T y = S(\lambda)$$

Again,  $(\log)\lambda$  is chosen with cross-validation.

# Descriptive Statistics

## Estimation of mean and covariance functions

Consider square integrable random curves

$$X = \{X(t), t \in [0, T]\}$$

and let  $X_1, X_2, \dots, X_m$  be an i.i.d. sample drawn from  $X$ .

**Mean function**  $\mu = E[X(t)]$  estimated by

$$\hat{\mu}(t) = \frac{1}{m} \sum_{i=1}^m X_i(t)$$

**Covariance function**  $c(t, s) = E[(X(t) - \mu(t))(X(s) - \mu(s))]$  estimated by

$$\hat{c}(t, s) = \frac{1}{m} \sum_{i=1}^m (X_i(t) - \hat{\mu}(t))(X_i(s) - \hat{\mu}(s))$$

## The Covariance operator

Associated with  $c(t, s)$  is the covariance operator  $C : L^2(T) \rightarrow L^2(T)$

$$f \mapsto C(f) : T \rightarrow \mathbb{R}$$

$$t \mapsto C(f)(t) = \int_T C(t, s)f(s)ds$$

estimated by

$$\hat{C}(x) = \frac{1}{m} \sum_{i=1}^m \langle X_i - \hat{\mu}, x \rangle (X_i - \hat{\mu})$$

Let  $\lambda_j$  and  $\psi_j$ ,  $j \geq 1$  be the eigenvalues and eigenfunctions of the covariance operator  $C$ :

$$C(\psi_j) = \lambda_j \psi_j \quad j \geq 1.$$

## Karhunen-Loève expansion

Assuming that  $\mathbf{E} \int_T (X(t) - \mu(t))^2 dt = \mathbf{E}(\|X - \mu\|^2) < \infty$  the random element  $X(t)$  can be expressed as

$$X(t) = \mu(t) + \sum_{j=1}^{\infty} Z_j \psi_j(t) \quad \text{where} \quad Z_j = \langle X - \mu, \psi_j \rangle \quad j \geq 1$$

This is known as the Karhunen-Loève expansion of  $X$ , and  $\{Z_j, j \geq 1\}$  is a sequence of centred random variables with  $\mathbf{E}(Z_j^2) = \lambda_j$ , and are uncorrelated, that is  $\mathbf{E}(Z_j Z_h) = 0$  for  $h \neq j$ .

Hence, when  $X$  is Gaussian, the  $Z_j$  are independent Gaussian random variables.

Moreover

$$\mathbf{E} \left( \int_T (X(t) - \mu(t))^2 dt \right) = \mathbf{E} \|X - \mu\|^2 = \sum_{j=1}^{\infty} \lambda_j$$

## Functional principal component analysis (FPCA)

Let  $\{X_n(t)\}$  be the sample of curves. We can understand FPC as the coordinates maximizing variability:  $p$  uncorrelated random variables  $Y_i = \langle X(t), v_i(t) \rangle$  for  $i = 1 \dots p$  such that

$$\text{Var} \left( \sum_{i=1}^p \langle X(t), v_i(t) \rangle \right)$$

is maximized subject to the constraint that  $|v_i| = 1$ .

It is possible to identify the principal components as the eigenfunctions with higher eigenvalues of the covariance operator  $C$  and build a data decomposition through the eigenvalues. We can also see that they are optimal orthonormal basis.

## Parenthesis: Principal components in $\mathbb{R}^p$

A principal component analysis (or PCA) is a way of simplifying a complex multivariate dataset. It helps to expose the underlying sources of variation in the data.

- Uses an **orthogonal transformation** to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables (PC)
- This transformation is defined in such a way that the first principal component has the largest possible variance and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set.
- PCA is sensitive to the relative scaling of the original variables.

## Examples in $\mathbb{R}^p$

<http://setosa.io/ev/principal-component-analysis/>

PCA can be easily performed in R with the command `prcomp`.

See the accompanying file `pc.R`

## Karhunen-Loève expansion and FPCA

- Assuming  $\lambda_1 > \lambda_2 > \dots > \lambda_j > \lambda_{j+1} > \dots$  and fixing an integer  $p$ , the orthonormal basis  $v_1, \dots, v_p$  solving the minimization problem

$$\min_{v_1, \dots, v_p} \mathbf{E} \left( \|X - \mu - \sum_{j=1}^p \langle X - \mu, v_j \rangle v_j\|^2 \right)$$

is formed by the first  $p$  eigenfunctions of the covariance operator,  $\psi_1, \dots, \psi_p$ , and the minimum is  $\sum_{j \geq p+1} \lambda_j$ .

- That is, the Karhunen-Loève expansion of  $X - \mu$  truncated at  $p$  gives the best  $p$ -dimensional approximation to  $X - \mu$  in the sense of mean square error.

$$X(t) \approx \mu(t) + \sum_{j=1}^p Z_j \psi_j(t) \quad t \in T$$

- The eigenfunctions  $\psi_j$  are also known as **principal functions**
- The  $Z_j = \langle X - \mu, \psi_j \rangle$  are known as **Scores** or **Functional Principal Components**.

## Sampling version of FPCA

The empirical version of the previous result is obtained by replacing the covariance operator  $C$  by the estimated covariance function

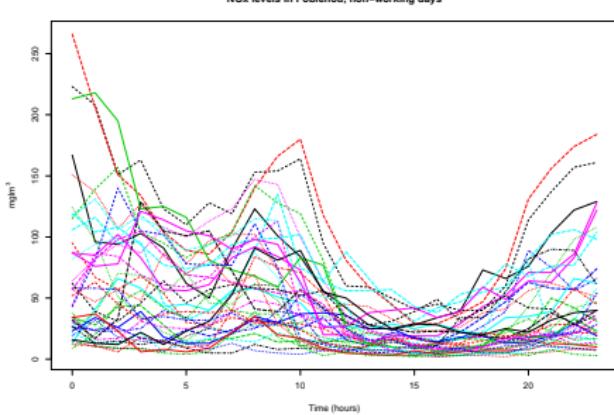
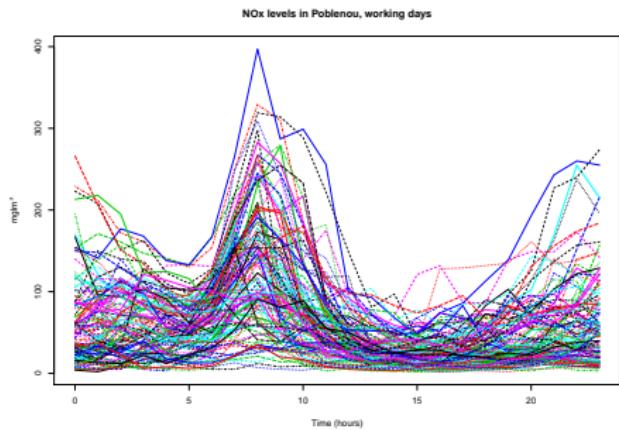
$$\hat{c}(t, s) = \frac{1}{n} \sum_{i=1}^n (X_i(t) - \bar{X}(t)) (X_i(s) - \bar{X}(s))$$

In practice, the eigensystem of the operator with kernel  $\hat{C}(t, s)$  is obtained by matrix diagonalization.

$$X_i(t) = \bar{X}(t) + \sum_{j=1}^p \hat{Z}_{i,j} \hat{\psi}_j(t) \quad t \in T, i = 1, \dots, n$$

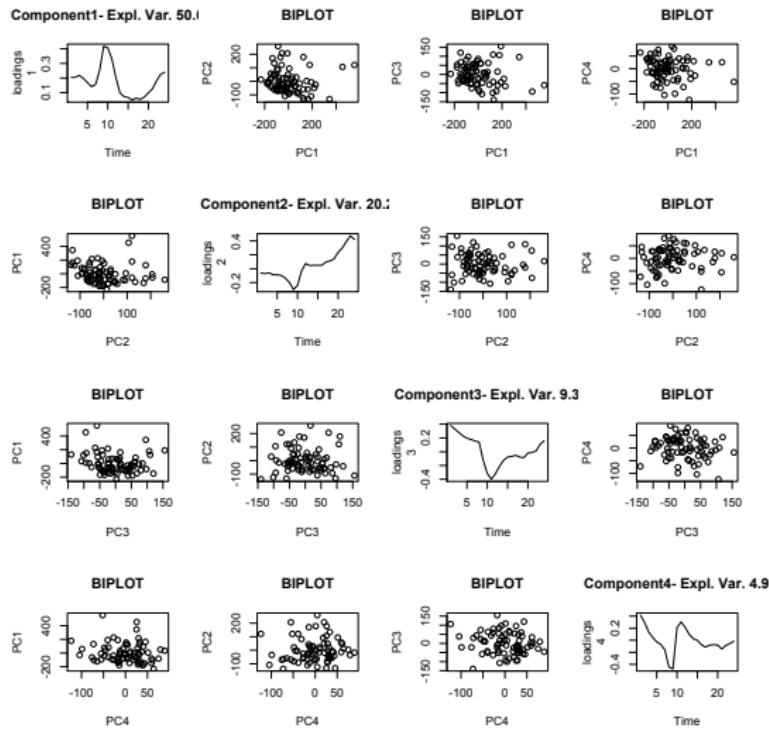
FPCA can be easily performed in R within the libraries `fda` and `fda.usc`.

# Example: NOx levels in Poblenou



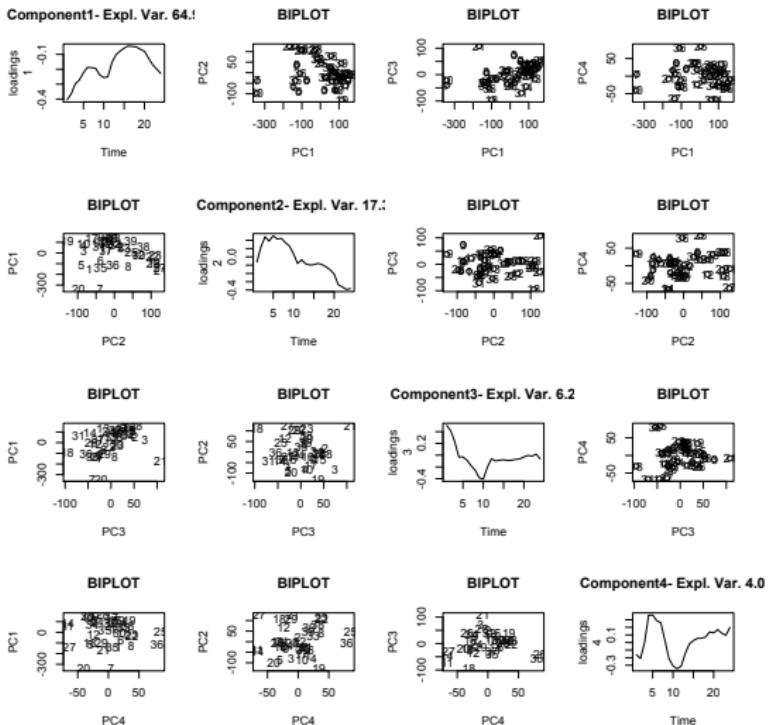
With 4 components 84.53 % of the variability in working-days is explained.

PC1	PC2	PC3	PC4
50.01	20.21	9.35	4.96



With 4 components 92.59 % of the variability of non-working days is explained.

PC1	PC2	PC3	PC4
64.92	17.32	6.28	4.08



# **Ordering: a first step towards inference**

# Depth Measures for Functional Data

Another way of approaching the functional data analysis is the introduction of the notion of depth, that is related to a generalisation of the concept of ordering for functional data.

The idea is to assign an order to each element of the sample related to its centrality for the whole set.

There are many measures of depth, and there is no agreement about their advantages.<sup>2</sup>

We will concentrate on the Fraiman-Muñiz depth, but before that, let's review other notions of depth.

---

<sup>2</sup>a good review is done in Estrada's Thesis, ULA (2014)

## The Fraiman-Muñiz Depth<sup>3</sup>.

Consider first a univariate sample

$$U_1, \dots, U_n \quad \text{and} \quad U_{(1)}, \dots, U_{(n)}$$

be the corresponding order statistics. For no ties, we have that  $U_i = U_{(j)}$  and then the natural depth of  $U_i$  is said to be

$$D_n(U_i) = \frac{1}{2} - \left| \frac{1}{2} - \left( \frac{j}{n} - \frac{1}{2n} \right) \right|$$

This notion of depth assigns minimal and equal depth to the two extreme values of the sample, maximum to the innermost point(s) and changes linearly with the position the datum occupies in the sample.

---

<sup>3</sup>Fraiman & Muñiz (2001) Trimmed means for functional data, *TEST* 10.

For the case of functional data, consider the sample  $X=\{X_i(t)\}$  defined in a common interval  $\mathcal{J}$ .

For each  $t$  we can compute the natural depth,  $D_n(X_i(t))$ , and then the depth for each  $X_i(t)$  is:

$$I(X_i, X) = \int_{\mathcal{J}} D_n(X_i(t)) dt$$

Where, in practice, the integral is replaced by a sum over the  $t$  for the time grid.

The deeper curve will be the  $X_i$  that with the biggest  $I(X_i, X)$ .

Defining a Wilcoxon test based on this ordering is a natural option for the 2-sample problem<sup>4</sup>

---

<sup>4</sup>Lopez-Pintado & Romo (2007), Depth-based inference for functional data, *CSDA* 51 suggest the use of this procedure with their band-depth.

## Some other depth Measures for Functional Data

- Band depth<sup>5</sup>

- ① The band determined by the curves  $X_{i_1}, \dots, X_{i_k}$  is

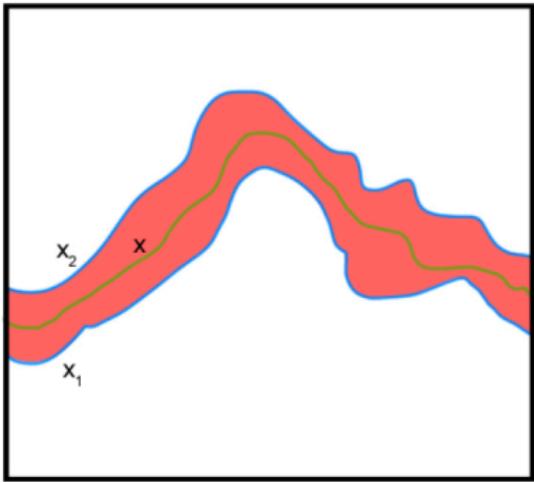
$$B_{i_1, \dots, i_k} = \{(t, y), \min_{1 \leq r \leq k} X_{i_r} \leq y \leq \max_{1 \leq r \leq k} X_{i_r}\}$$

Define  $S_i^j$  as the proportion of the bands formed by  $j$  curves that contain the graph of  $X_i(t)$ , and the depth of  $X_i$  as the sum  $\sum_{j=1}^J S_i^j$  for a fixed  $J$ .

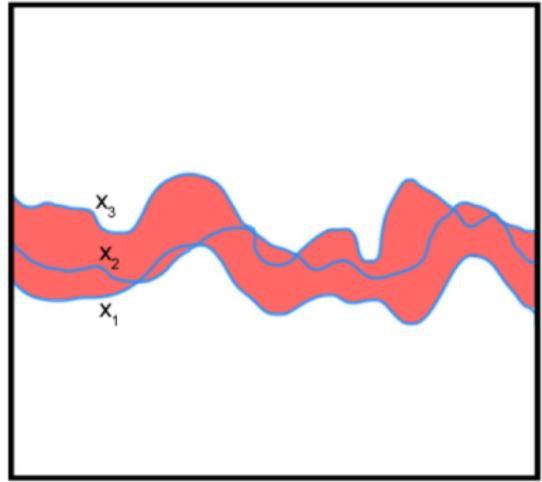
- ② Generalize  $GS_i^j$  considering the amount of time (Lebesgue measure of the set) that the curve  $X_i$  is inside the band, and again the sum of these will be the depth.

---

<sup>5</sup>Lopez-Pintado & Romo (2009), On the concept of depth for functional data JASA 104



$x \in$  band defined by  $x_1$  and  $x_2$



Band defined by  $x_1$ ,  $x_2$ ,  $x_3$

## Some other depth Measures for Functional Data

- *h-modes*<sup>6</sup>

For each curve  $X_i$ , its *h-mode* is defined as

$$\hat{f}_h(X_i) = \frac{1}{n} \sum_{j=1}^n K_h(||X_j - X_i||)$$

where  $K_h(t) = 1/hK(t/h)$  and  $K$  is the Gaussian kernel.

The deepest curve will be that for which  $\hat{f}_h$  is maximum. The choice of  $h$  is arbitrary...

---

<sup>6</sup>Cuevas, Febrero, Fraiman (2007), Robust estimation and classification for functional data via projection-based depth notions, *Computational Statistics* 22

## Some other depth Measures for Functional Data

### Projection over principal components<sup>7</sup>

Project  $X_i$  on the first  $h$  principal components, and take a weighted average of the univariate depth of the projection points.

### Random Projections<sup>8</sup>

Given a random direction  $a$ , project the curve  $\gamma$  on  $a$  (this requires some regularity condition to ensure that the inner product is well defined), and consider the univariate depth of the projected points.

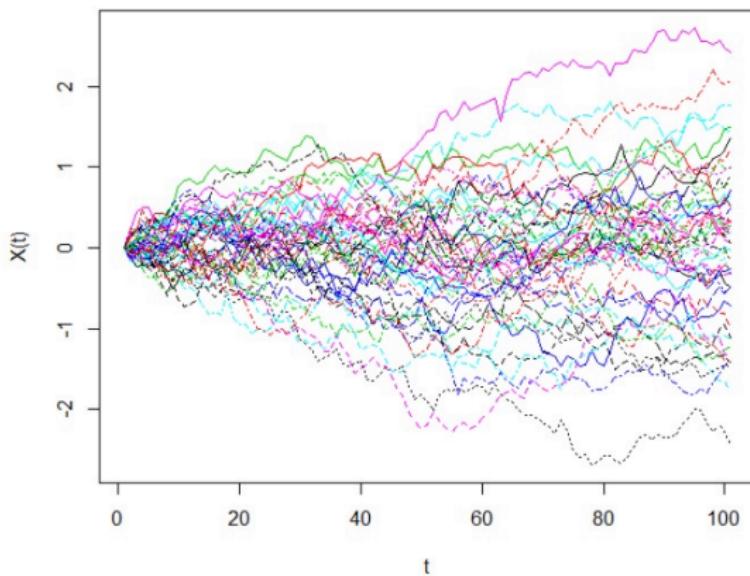
The depth of the curve is then defined as the average of the univariate depths of several random projections.

---

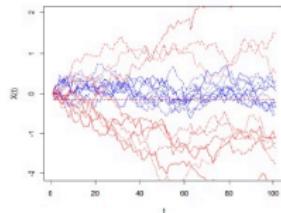
<sup>7</sup>Estrada, Quiroz (2014)

<sup>8</sup>Cuevas, Febrero, Fraiman (2007), *op.cit.*

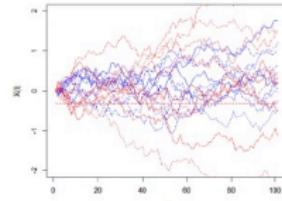
## Example: 50 curves from Brownian Motion



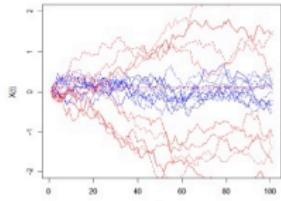
# Deepest and shallowest curves



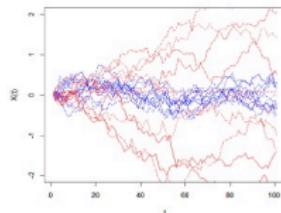
(a) Fraiman y Muñiz



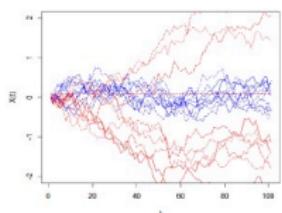
(b) Prof. por bandas (V1)



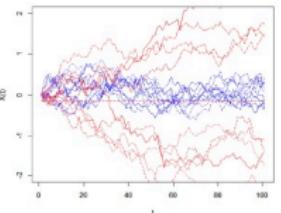
(c) Prof. por bandas (V2)



(d) h-moda



(e) Proyección aleatoria

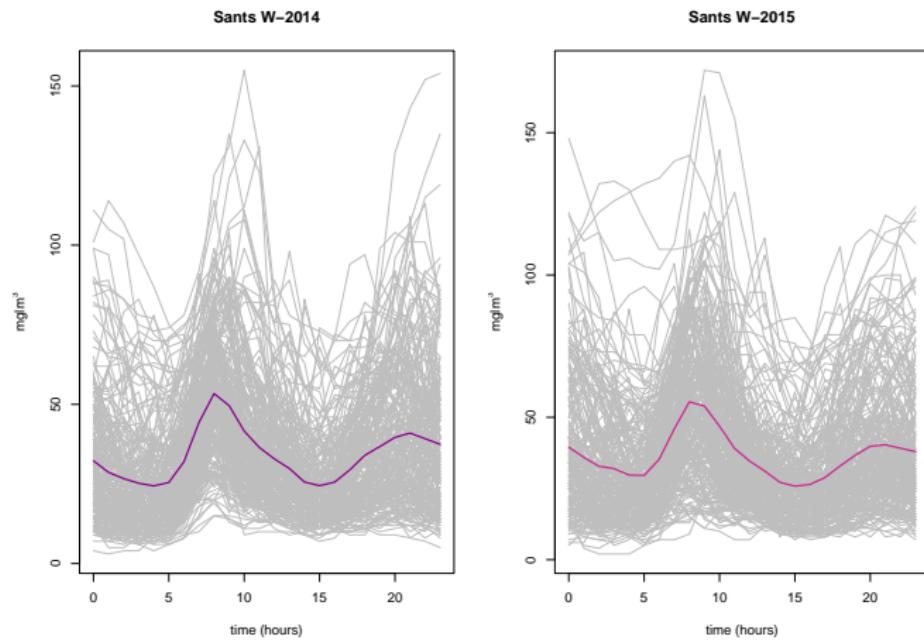


(f) Proyección PCA

Bands:  $J = 2$ , Modes:  $h = 0.1$ , 2 principal components (83.74% & 7.75 % resp.), and 50 random projections.

# An example of Inference: The two-sample problem

$\text{NO}_2$  levels on working days taken on the same station, in 2014 and 2015. Did (the law) of contamination- levels change?



## On Permutation tests

Permutation tests are one of the most important tools of non-parametric statistical methodology.

A permutation test gives a simple way to compute the sampling distribution for any test statistic, when the distribution is invariant under permutations for the null hypothesis.

To estimate the sampling distribution of the test statistic we randomly shuffle the exposures to make up a large enough sample of data sets.

If the null hypothesis is true, this shuffled data sets should have all the same distribution, and thus should look like the original data, otherwise they should look different.

## Obtaining $p$ -values

Let  $N = m + n$  and  $Z_1, \dots, Z_N$  be the combined sample obtained by concatenating the  $X$  and  $Y$  samples. Relevant  $p$ -values for the statistics we shall consider here can be obtained as follows:

- ① *From the joint sample,  $Z$ , select a random subset of size  $m$ . Declare that the chosen elements belong to the  $X$  sample and the remaining  $n$  to the  $Y$  sample.*
- ② *Compute the statistic of interest for this artificial pair of samples.*
- ③ *Repeat steps (1) and (2) a large number  $B$  of times, and*
- ④ *From the  $B$  values computed above, extract an approximate  $p$ -value for the observed statistic (the one calculated with the original  $X$  and  $Y$  samples)*

## Two sample problem for functional data

Let  $X_1(t), \dots, X_m(t)$  denote an i.i.d. sample of real valued curves defined on some interval  $J$ .

Let  $\mathcal{L}(X)$  be the common probability law of these curves.

Likewise, let  $Y_1(t), \dots, Y_n(t)$ , be another i.i.d. sample of curves, independent of the  $X$  sample and also defined on  $J$ , with probability law  $\mathcal{L}(Y)$ .

We want to test the null hypothesis,

$$H_0 : \mathcal{L}(X) = \mathcal{L}(Y) \text{ against the general alternative } \mathcal{L}(X) \neq \mathcal{L}(Y)$$

We shall discuss different permutation tests<sup>9</sup>:

- A functional Shilling test
- A Wilcoxon type test
- Another test based on depths, that uses meta analysis ideas to assess significance
- This last test can be re-defined in terms of FDRs (... I think, anyone wants to give it a thought?).

and compare their performance with the classical test by Kokoszka and Horváth, based on the principal components of the pooled covariance operator of the two samples.

---

<sup>9</sup>A. Cabaña, et.al. Permutation tests in the two-sample problem for functional data, in *Functional Statistics and related Fields*, G. Aneiros, E. Bongiorno, R. Cao & P. Vieu, Ed. Springer, (2017)

## Schilling test

This is a generalisation to functional data of the  $k$ -nearest neighbours multivariate two-sample test of Schilling<sup>10</sup>.

Let  $N = m + n$  and  $Z_1, \dots, Z_N$  be the combined sample obtained by concatenating the  $X$  and  $Y$  samples.

Define the indicator function

$$I_i(r) = \begin{cases} 1 & \text{if } Z_i \text{ and its } r\text{-nn are from the same sample} \\ 0 & \text{else} \end{cases}$$

Nearest neighbours are based on  $L^2$  distance, and with probability 1 they are uniquely defined. In case of ties, we would decide at random.

---

<sup>10</sup>Schilling (1986) Multivariate two-sample tests based on nearest neighbors, *JASA* 81, 799-806

In practice, if the functions have been registered in a common grid, say  $0 = t_0 < t_1 < \dots t_L = T$ , a reasonable approximation to the distance between functions  $Z_i$  and  $Z_j$  would be

$$d_{i,j} = \sum_{l=1}^L \Delta_l (Z_i(t_l) - Z_j(t_l))^2, \text{ where } \Delta_l = t_l - t_{l-1}$$

If the grid used is equally spaced,  $\Delta_l$  can be omitted and the curves can be treated as points in  $\mathbf{R}^L$  in order to compute faster the  $k$ -nearest-neighbours.

When no common grid is available, represent the functions in the joint sample in terms of local polynomials, or some other basis functions, and the  $k$ -nearest-neighbours are identified by a quadratic algorithm (in the joint sample size  $N$ ).

Define the statistic:

$$T_{N,k} = \frac{1}{Nk} \sum_{i=1}^N \sum_{r=1}^k I_i(r)$$

Under  $H_0$  we expect  $T_{N,k}$  to be small.

Observe that the expected value

$$\mathbf{E} T_{N,k} = \mathbf{E} I_i(r) = \frac{m(m-1) + n(n-1)}{N(N-1)}$$

while the variance depends on the amount of pairs of points that are mutual neighbours and the amount of pairs that share a common neighbour.

In this case, the belonging of one element to either  $\{X_m(t)\}$  or  $\{Y_n(t)\}$  is an arbitrary choice, for they represent just two subgroups of a larger sample generated from a single random experiment. Thus, we can estimate the distribution of the statistic through permutations on the labelling.

## Algorithm

- ① For the concatenation of the samples  $Z$ , keeping the natural ordering, we compute the  $m \times n$  matrix of distances between its elements.
- ② Consider we are running the test for  $k$  neighbours. We build a  $N \times k$  matrix that in the  $i$ -th row contains the indices of the  $k$  nearest elements to the curve  $Z_i$ .
- ③ In order to compute the statistic it is enough to count how many of the indices in each of the first  $m$  rows is equal or less than  $m$ , (i.e. are originally  $X$ ) and how many of the indices corresponding to  $m + 1, \dots, N$  is greater than  $m$ .
- ④ Obtain the distribution of  $T_{N,k}$  using the permutation procedure.

In order to compute the distribution using the permutation procedure, we do the following:

- To the elements of the original nearest neighbour table, apply a random permutation  $\tau$ .
- In the  $r$ -th row, if  $\tau(r) \leq m$ , count the number of elements that are less than or equal to  $m$ , else, count the number of elements greater than  $m$ .
- with the numbers obtained in (ii) compute  $T_{N,k}$ .

All these operations can be performed in time  $O(Nk)$ , thus the computational cost of an iteration is basically linear in  $N$ . As for the initial cost of setting the neighbours indices table, several sub-quadratic algorithms have been developed for this problem, since the fundamental contribution of Friedman *et.al.*<sup>11</sup>.

---

<sup>11</sup>Friedman, Baskett & Shustek (1975) An algorithm for finding nearest neighbours, *IEEE Trans. Com.* 24

## Back to the 2-sample problem: Meta-analysis based tests

Let  $p_i$  be an empirical  $p$ -value related to the *centrality* of the variable  $Y_i$  on the sample  $X$ ,

$$\hat{p}_* = \frac{\sum_{j=1}^m I(D(X_j|X) \leq D(Y_*|X))}{m+1} \quad \begin{cases} p_i = \hat{p}_i & \text{if } \hat{p}_i \neq 0 \\ p_i = \frac{1}{m+1} & \text{if } \hat{p}_i = 0 \end{cases}$$

Under  $H_0$  (ignoring ties), each  $p_i$  is uniformly distributed in  $\{1/m, 2/m, \dots, 1\}$ , but they are not independent.

For symmetry, we consider as well the  $q_i$   $p$ -values related to the *centrality* of the variable  $X_i$  on the sample  $Y$ .

$$\hat{q}_* = \frac{\sum_{j=1}^n I(D(Y_j|Y) \leq D(X_*|Y))}{n+1} \quad \begin{cases} q_i = \hat{q}_i & \text{if } \hat{q}_i \neq 0 \\ q_i = \frac{1}{n+1} & \text{if } \hat{q}_i = 0 \end{cases}$$

Observe that when  $H_0$  does not hold, the depth of a curve in a family where it does not belong will be very low, and so would be its associated  $p$ -value. In that case the corresponding statistic will be very big.

Consider<sup>12</sup>

$$S_Y = \sum_{i=1}^m -\ln(p_i) \quad S_X = \sum_{i=1}^n -\ln(q_i)$$

We want to associate a  $p$ -value to the pair  $(S_X, S_Y)$ .

---

<sup>12</sup>Hedges & Olkin (1985), *Statistical Methods for Meta-Analysis* Academic Press

When the two samples display a difference in “scale”, most, of the curves of the (say)  $X$  sample, turn out to be central with respect to the  $Y$  sample and  $S_X$  will not show a significant value. In such a situation, typically, several curves of the  $Y$  sample will turn out to be clearly outlying respect to the  $X$  sample, and the maximum will reach a significant value.

Hence, one possibility is considering

$$S = \max\{S_X, S_Y\}$$

In order to assess significance to the observed value of  $S$ , apply the permutation procedure.

Better empirical performance is obtained using following result:

**Lemma: Combining  $p$ -values**

Let  $p_X$  ( $p_Y$ ) denote the  $p$ -value of  $S_X$  ( $S_Y$ ), under the null permutation distribution, as obtained from procedure  $p$ -value if all subsets of size  $m$  were used (instead of just a sample of size  $B$ ) and assuming the null hypothesis. Then

- ①  $\Pr(p_X \leq t) \leq t$  for any  $t \in (0, 1)$ , and the same holds for  $p_Y$ .
- ②  $\Pr(2 \min(p_X, p_Y) \leq t) \leq t$  for any  $t \in (0, 1)$ .

The null permutation distribution of  $S_X$  is a discrete distribution that can not be assumed uniform on its range (some values of  $S_X$  can appear more frequently than others when subsets are chosen at random). This is why (1) is not obvious.

Part (2) of the Lemma tells us that an appropriate  $p$ -value for  $2 \min(p_X, p_Y)$  is the observed value of this statistic itself.

Thus, our second way of getting a  $p$ -value from  $S_X$  and  $S_Y$  is the following:

Compute, approximately,  $p_X$  and  $p_Y$  for  $S_X$  and  $S_Y$ , respectively, using the procedure  $p$ -value described and use  $2 \min(p_X, p_Y)$  as  $p$ -value.

In the experiments described next,  $p_X$  and  $p_Y$  was computed using independent “draws” of subsets of the joint sample, a procedure that yields good power results.

## Empirical comparison of powers

In order to fix a standard, we have also computed the empirical power for Horvath and Kokoszka's test for equality of mean functions. The null hypothesis is rejected for large values of

$$U_{m,n} = \frac{mn}{m+n} \int (\bar{X}_m(t) - \bar{Y}_n(t))^2 dt$$

Under some regularity conditions, the asymptotic distribution of  $\bar{X}_m(t) - \bar{Y}_n(t)$  is a Gaussian process  $\Gamma$  whose covariance can be approximated by the pooled covariance operator of the two samples, hence, the distribution of  $U_{m,n}$  can be approximated by the first  $d$  terms in the Karhunen-Loève expansion of  $\int_0^1 \Gamma^2(s) ds \approx \sum_{i=1}^d \lambda_i N_i^2$ ,  $\lambda_i$  are the (ordered) eigenvalues of the pooled covariance estimator, and  $N_i$  are i.i.d.  $N(0,1)$ .

## A simulation experiment

We have simulated samples of functional data as realisations from a geometric Brownian motion process

$$f(t) = X_0 \exp \left\{ rt - \frac{t\sigma^2}{2} + \sigma w_t \right\}$$

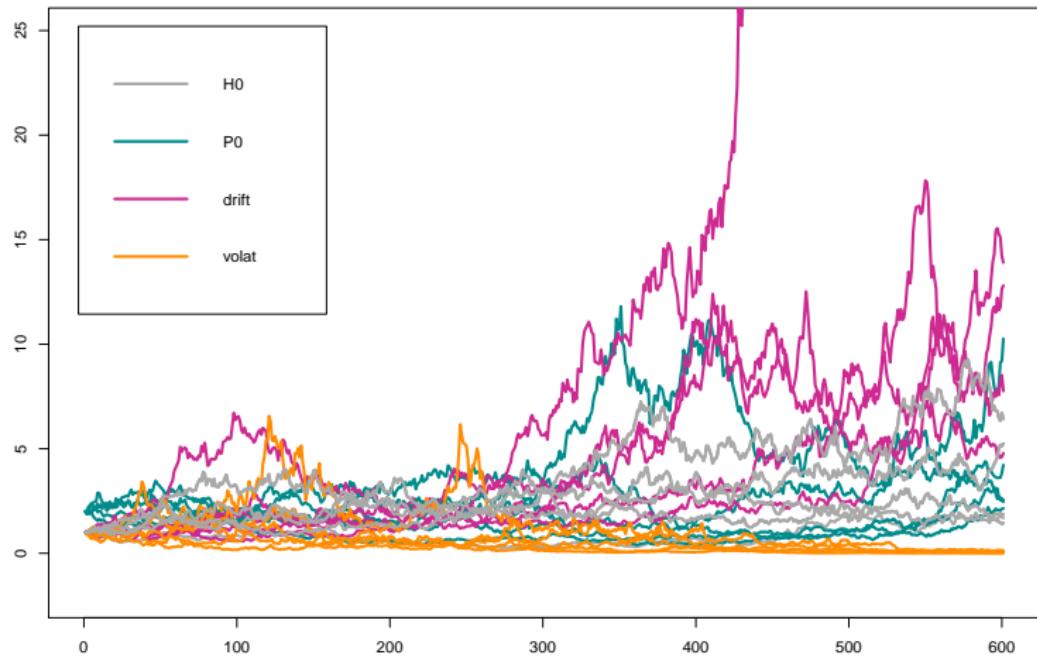
where  $r$  and  $\sigma$  are, respectively, the trend (drift) and volatility coefficients,  $w_t$  is a standard Wiener process and  $X_0$  is the initial value.

We have tested

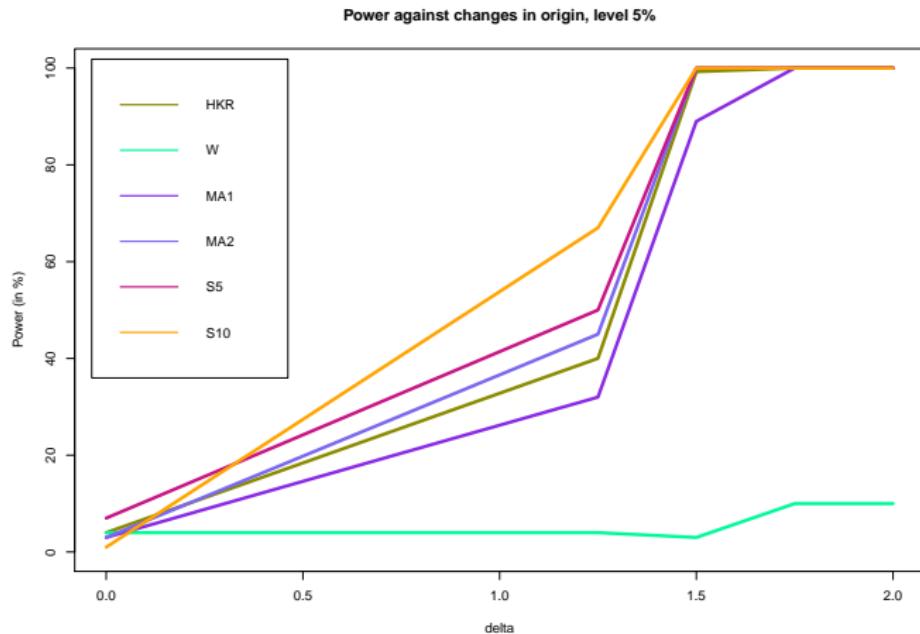
$$H_0 : \mathcal{L}(X) = \mathcal{L}(Y), \quad \text{against} \quad \mathcal{L}(X) \neq \mathcal{L}(Y)$$

where  $X$  was simulated with  $r = 1, \sigma = 1, X_0 = 0$  and  $Y$  is any of the 'contaminated' samples with only one of the parameters varying at a time.

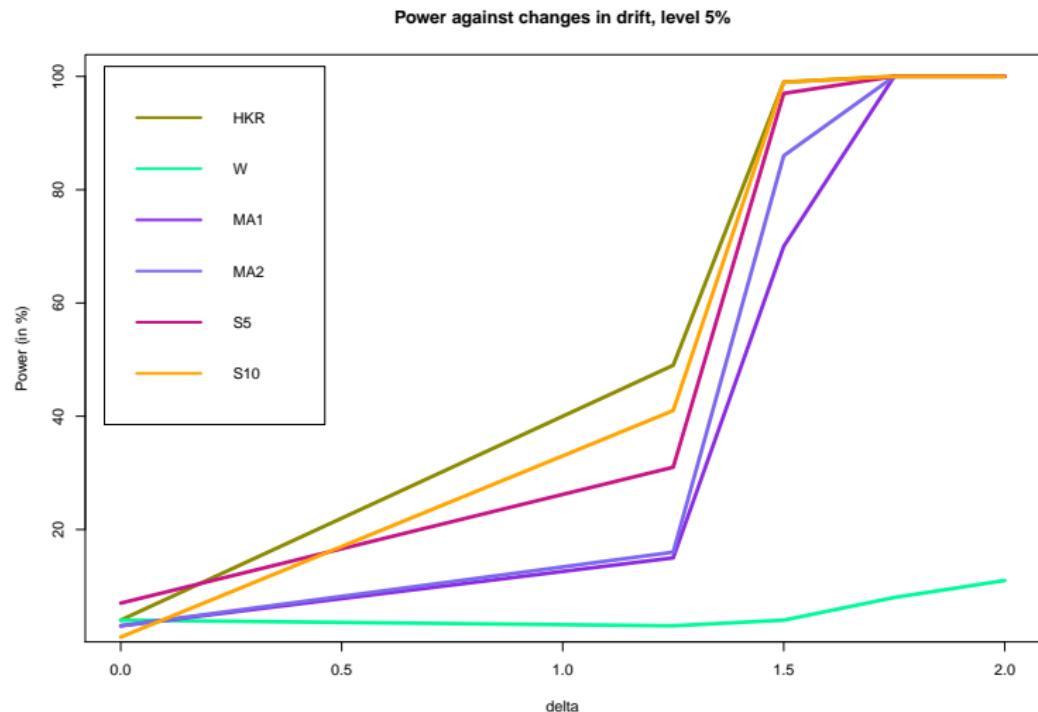
# A few curves



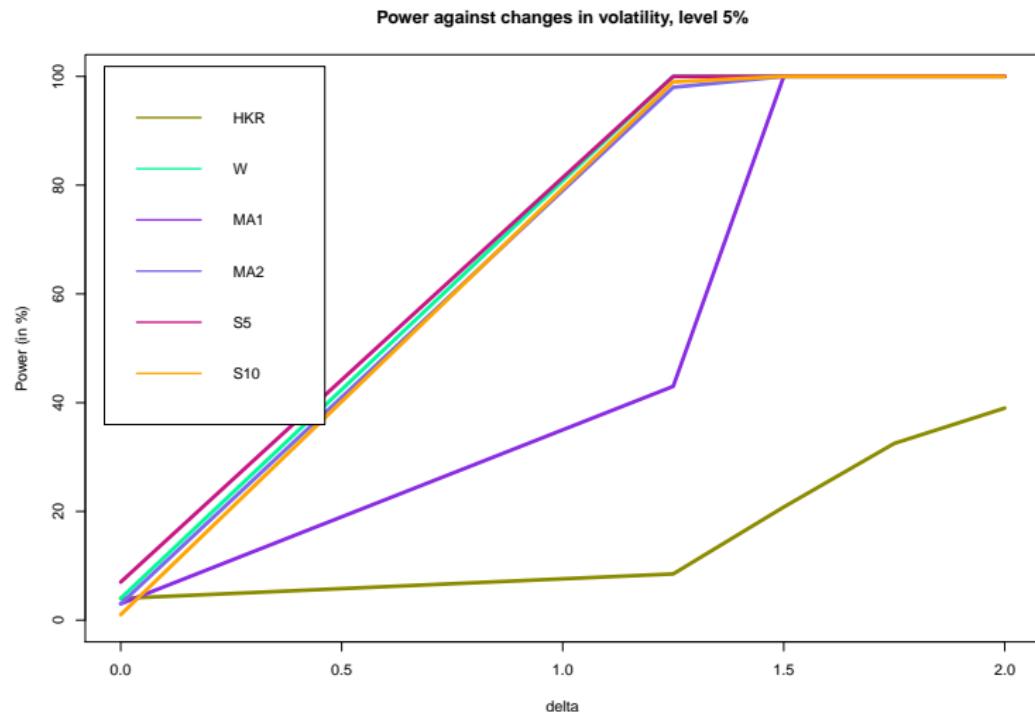
# Empirical power of tests, Geometric Brownian motion



# Empirical power of tests, Geometric Brownian motion



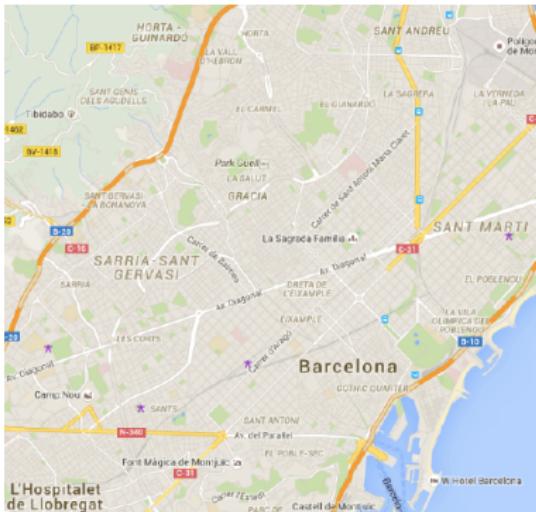
# Empirical power of tests, Geometric Brownian motion



# Comparing the NO<sub>2</sub> levels in different neighbourhoods

We have hourly measurements of nitrogen dioxide, a known pollutant, in four neighbourhoods in Barcelona.

The measurements were taken along the years 2014 and 2015 in automatic monitoring stations<sup>13</sup>



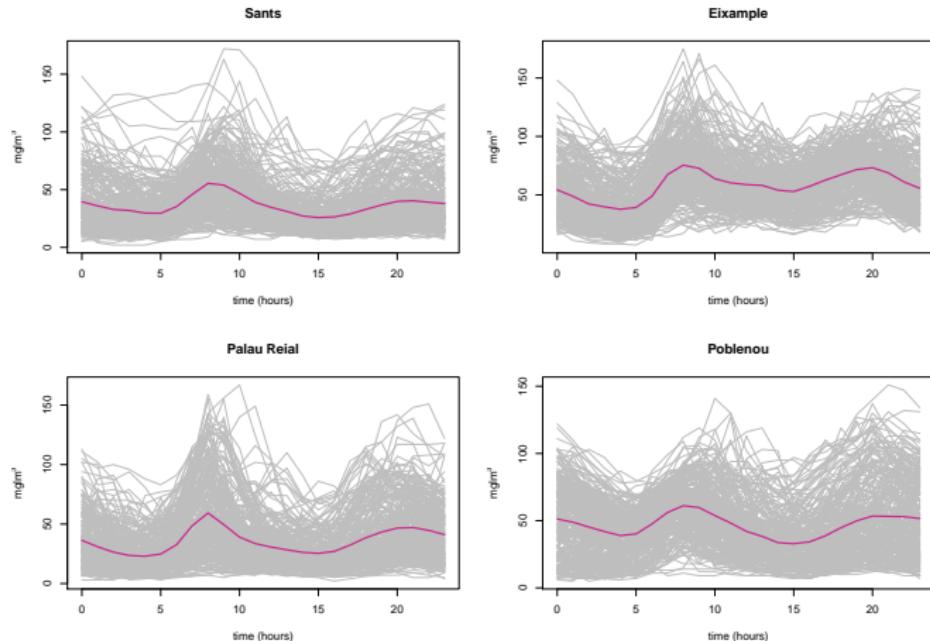
<sup>13</sup>available from <http://dtes.gencat.cat/icqa>.

We have split the data sets in workingdays ( $\approx 220$  curves) and non-working days ( $\approx 120$  curves), each year.

Questions of interest are

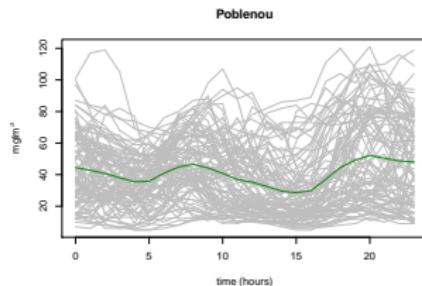
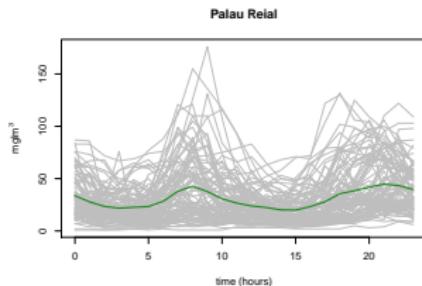
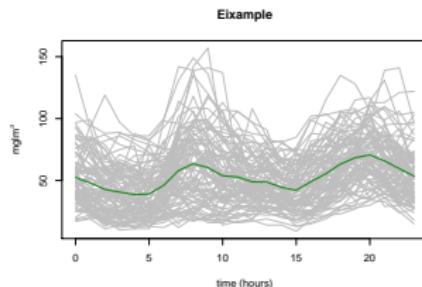
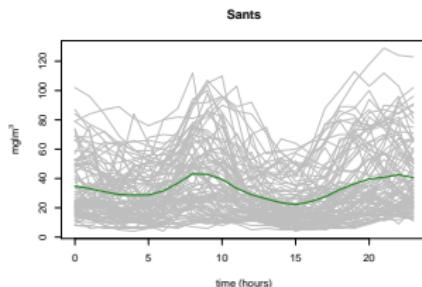
- Is the level of polutants significantly different during working and non-working days?
- Did the levels of  $\text{NO}_2$  changed from one year to the next in each of the neighbourhoods?
- Comparing the pollution levels among the different neighbourhoods.

# $\text{NO}_2$ levels on working days, 2015



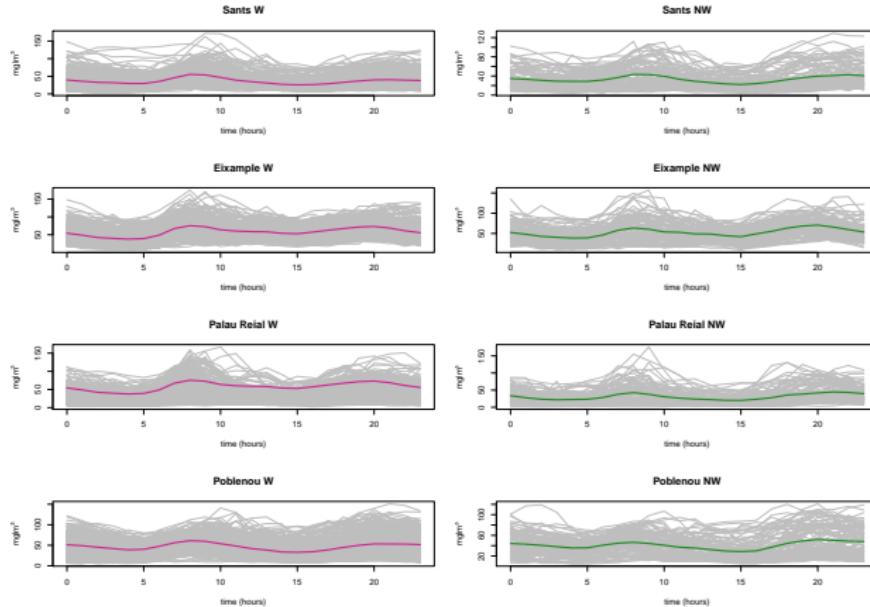
Schilling 10 and HK show marginal or no differences between pairs of neighbourhoods, the other tests show low p-values, indicating pairwise differences.

# $\text{NO}_2$ levels on non- working days, 2015

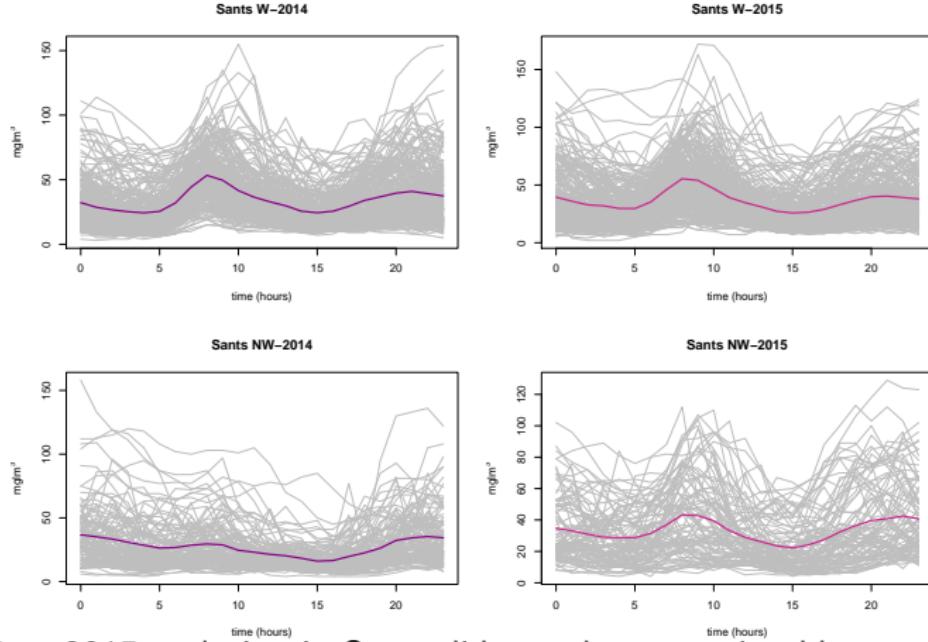


Moderate  $p$ -values ( $\approx 0.08$ ) are found for Sants-Poblenou in 2015, all other pairs are significantly different.

# $\text{NO}_2$ levels on working and non- working days, 2015



All tests show evidence of differences between working and non-working days, with Wilcoxon and Schilling-10 showing the strongest evidence of differences.



From 2014 to 2015, pollution in Sants did not change noticeably on non-working days, but significant changes are found from one year to the next, on working days, with the Wilcoxon and Schilling procedures being the ones that find stronger evidence of change.