Course “Data Analysis” ModIA double degree INSA Toulouse / ENSEEIHT

Olivier Roustant

2023-09

General purpose

The aim of the course is

to analyze vector data

Remarks:

- more complex data (signal, images, text, etc.) almost always come down to vector data.
- understanding data is helpful (if not necessary) for modeling, and for interpreting model outputs.

Course functioning

To reach the aim, the course will be composed of:

- Courses on new material
- Computer labs on these materials, supported by case studies
  - **Run/complete** tutorials written in R and Python
  - **Answer** additional questions about interpretation
- Computer labs on a realistic problem $\rightarrow$ "class project"
  - **Write YOUR** tutorial in R
  - **Explain YOUR** methodology and interpretation

Evaluation

- Two individual exams, without documents (1/3 of the grade each)
  - → To test your knowledge on the program
  - → Questions on both theory and practice (interpretation)
  - → Timing. Exam #1: middle of period 2, Exam #2: end of period 3.
- The class project, by groups of 3 (1/3 of the grade)
  - → To test your know-how on a realistic case-study
  - → Deliverables: professionnal report + R code

Resource and program

Wikistat

The main resource of the course is **wikistat** (thanks Philippe Besse!), seasons 2 & 3, which is a great amount of information and experience!

Program

1. Principal component analysis and applications
2. Clustering: k-means, hierarchical clustering, Gaussian mixtures

Remarks:

- During labs: practice of R and Python, descriptive statistics
- Linear methods are taught in parallel in another UF.

A brief analysis of data analysis

There are a huge number of problems on data analysis.
Fortunately, there are common points between them!

- A few *classes* of engineering questions
  - $\rightarrow$ prediction (a value, a class), anomaly detection
- A few *classes* of models / methods
  - $\rightarrow$ linear models, trees, random forest, aggregation, kernel methods (e.g. SVM), boosting, neural networks, ...
- A few number of traps
  - $\rightarrow$ overfitting, correlation is not causality, unfair learning

Simplified instructions for data analysis

Before using models, there are a few actions to do:

- **Explore the data**
  - look for outliers, visualize, understand
- **Create new relevant variables or "features"**
  - The most influential variables are not always given in the dataset
- **Adapt the data to fit with models!** (lazy but often successful!)
  - convert to numerics (algorithms use numbers)
  - reduce dimension (sampling, PCA, basis decomposition)
  - use transformations (logarithm?)
  - use embeddings (indicator variables, kernel tricks)
- **Split the data in learning set / test set**
  - necessary to avoid overfitting

Example 1. To predict short term electric consumption

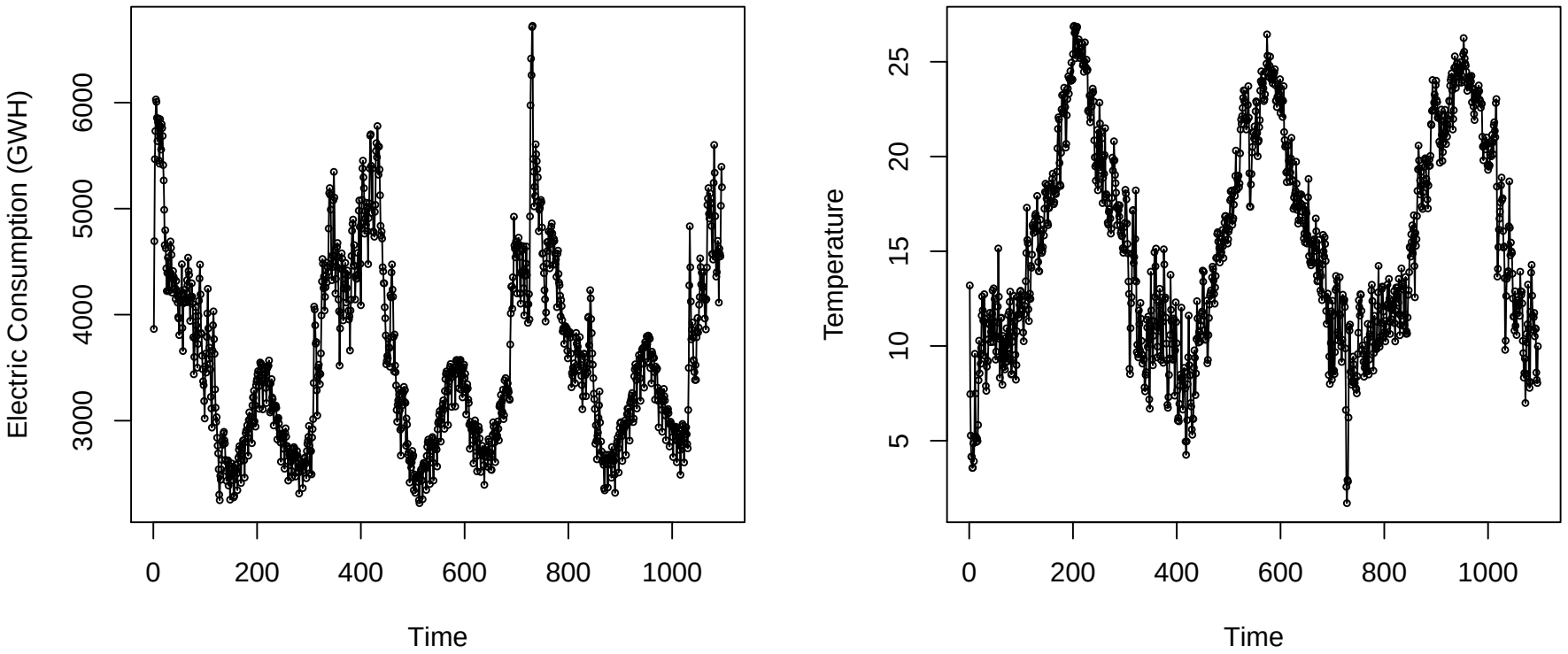


Figure: Daily electric consumption and temperature. Source: RTE and Meteo France

Example 1. To predict short term electric consumption

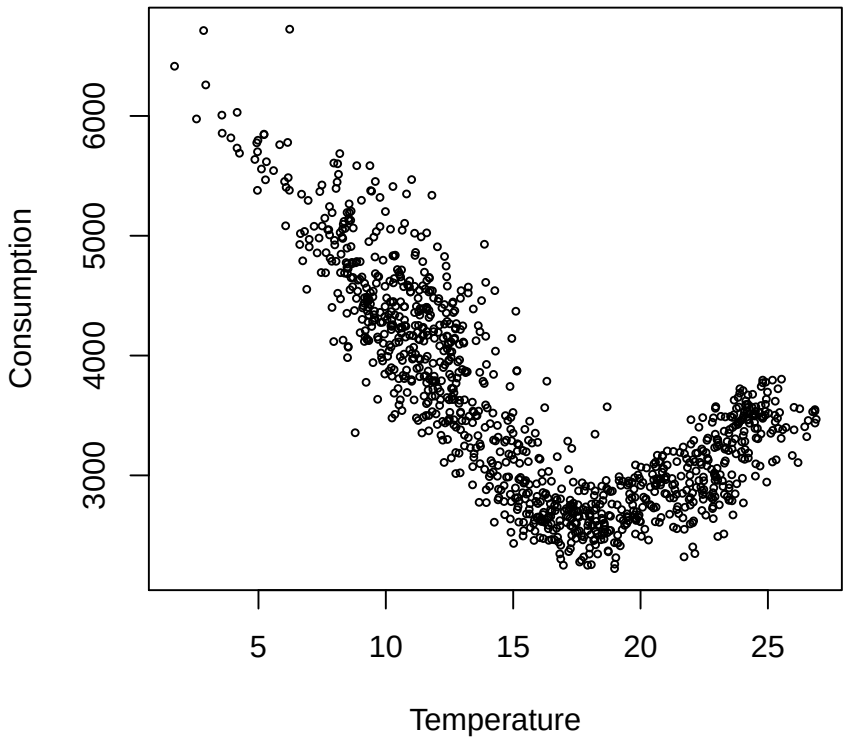


Figure: Electric consumption versus temperature.

Example 1. To predict short term electric consumption

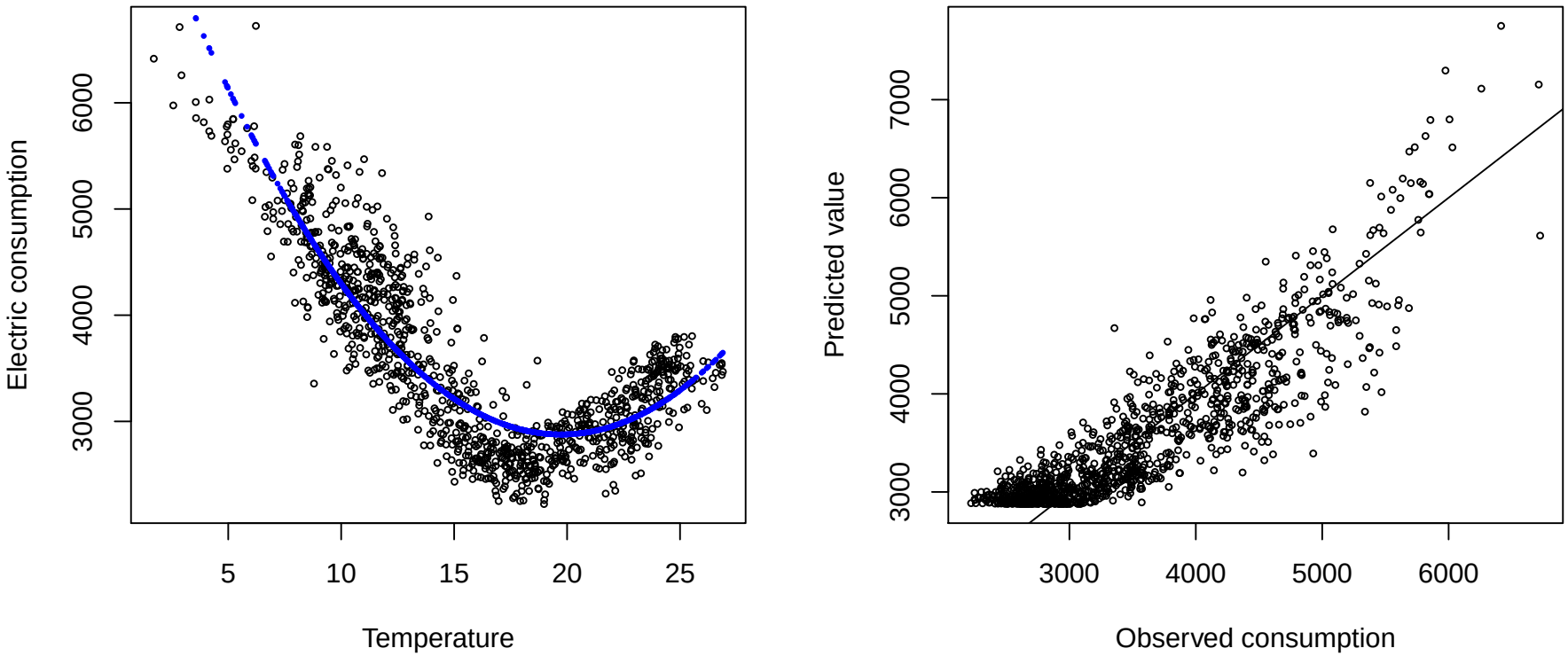


Figure: Linear models with only one predictor.

Example 1. To predict short term electric consumption

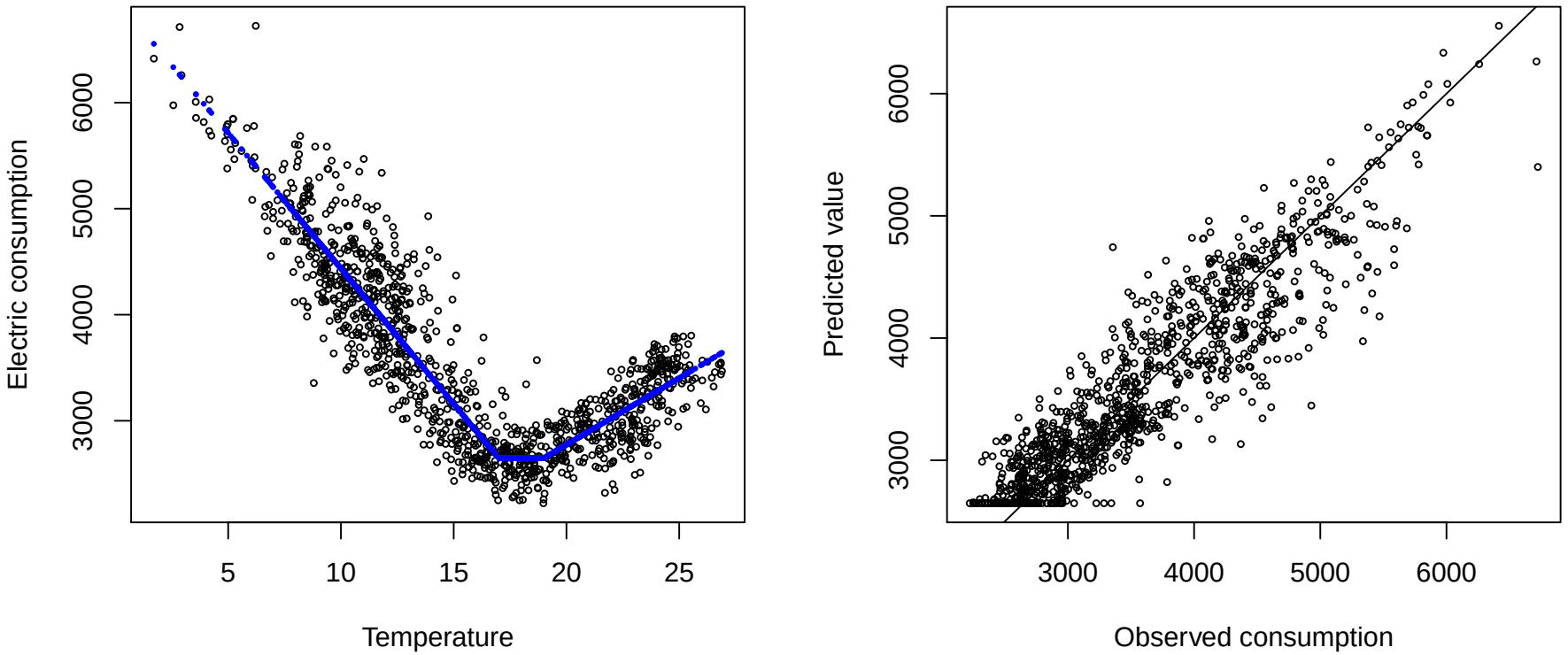


Figure: Linear models with only one predictor.

Example 1. To predict short term electric consumption

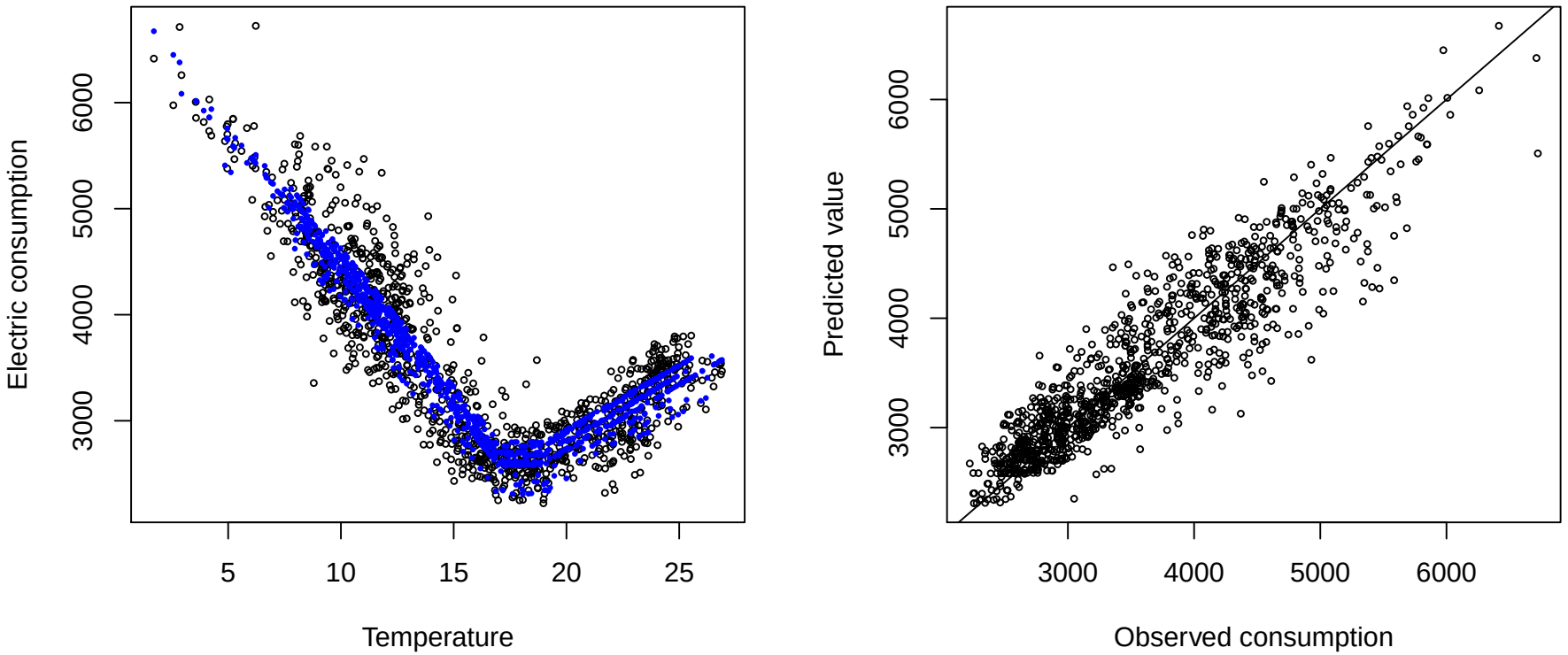


Figure: Linear model with new variables (e.g. day, off-day).

Example 1. To predict short term electric consumption

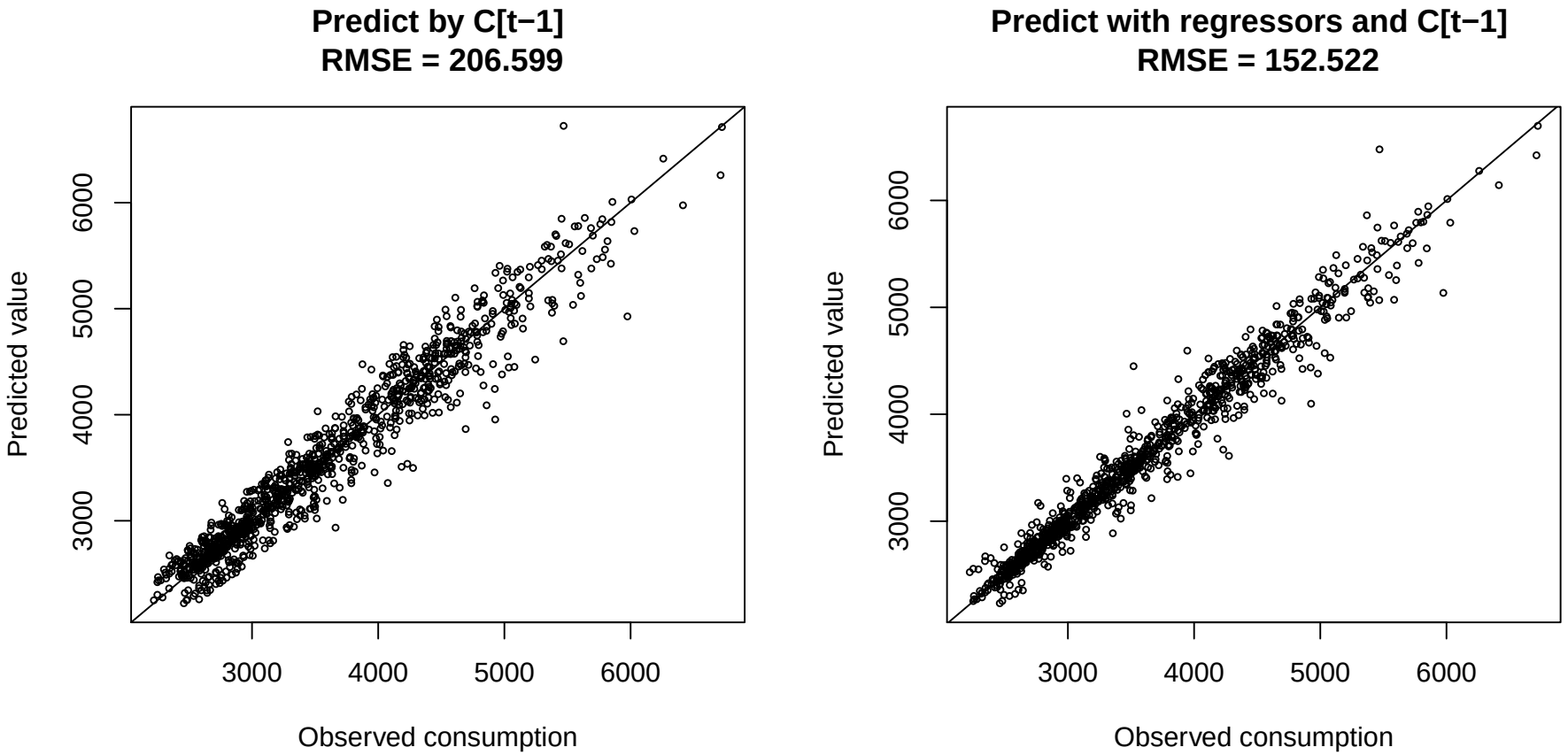


Figure: Linear model with new variables + lagged variable (ARMAX model).

Example 2. To detect anomalies in wafers

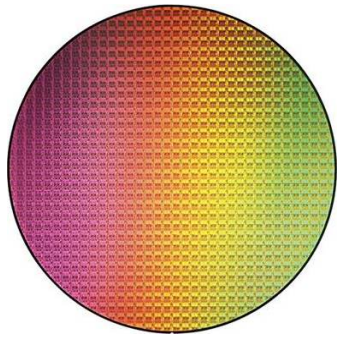
Connected objects

Integrated circuits

Industrial background

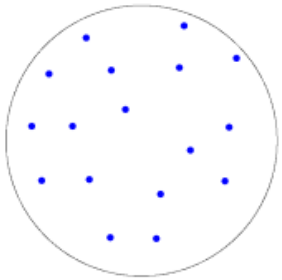
- PhD thesis of Espéran Padonou(*), with STMicroelectronics.
  ⇒ Study and production of integrated circuits.
  ⇒ Production on circular batches called wafers.
- High quality standards and costly control steps.
  ⇒ Probabilistic models for Advanced Process Control.

These slides are adapted from E. Padonou work



A silicon wafer

Example 2. To detect anomalies in wafers



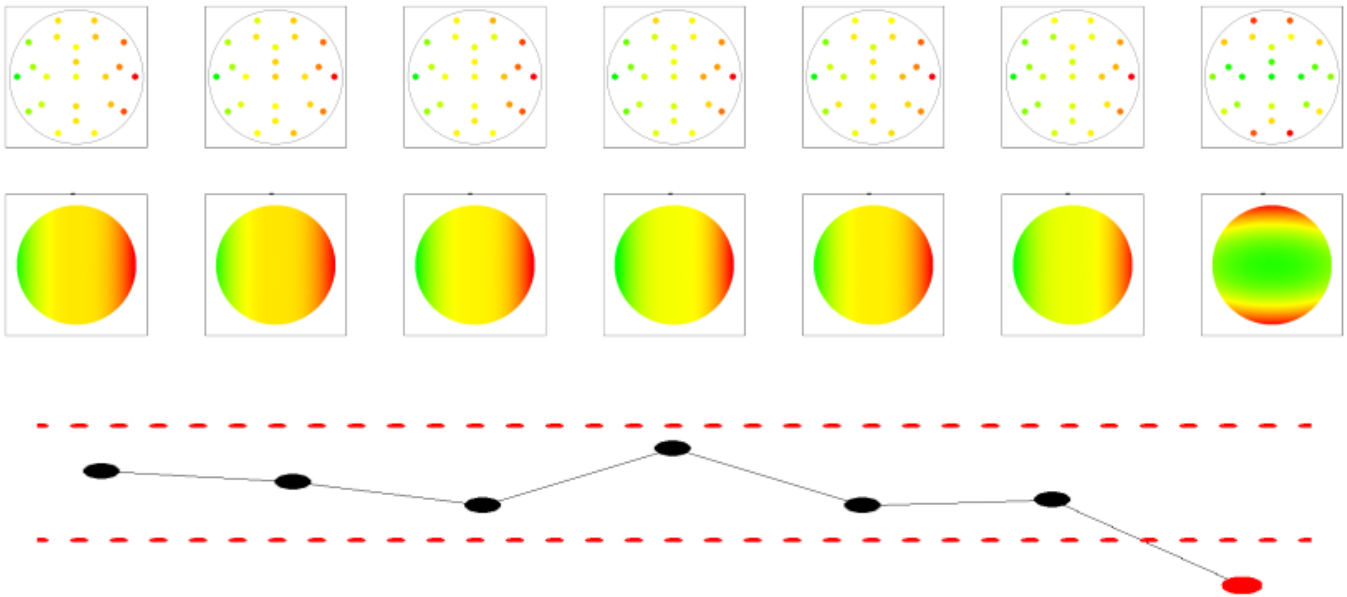
Profile Monitoring (Noorossana et al., 2011; Woodall, 2007)

1. At each time stamp, fit the model $y_i = f(\mathbf{x}^{(i)})$,
2. Monitor the parameters of f with standard control charts,

where $y_1, \ldots, y_n$ are measurements at the points $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$.

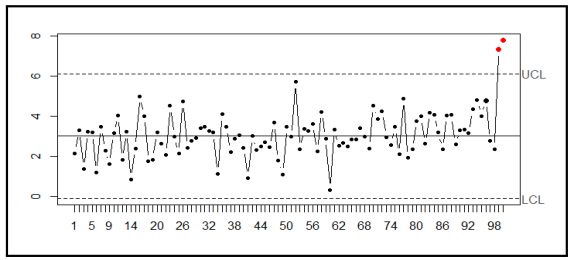
Designs of Experiments | Response surface models | Statistical Process Control

- DoE
- RSM
- SPC

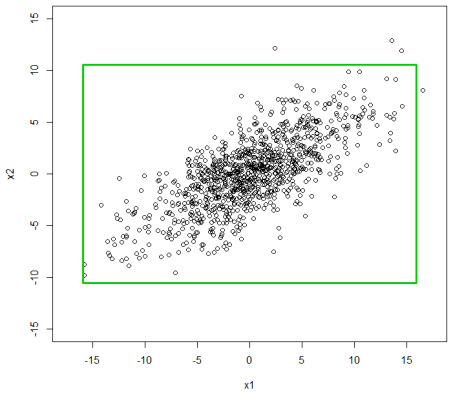


Example 2. To detect anomalies in wafers

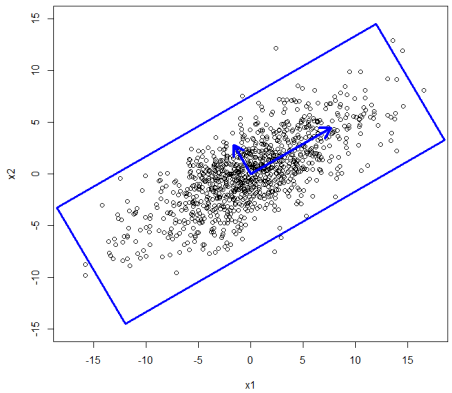
Examples of control charts
in advanced process control



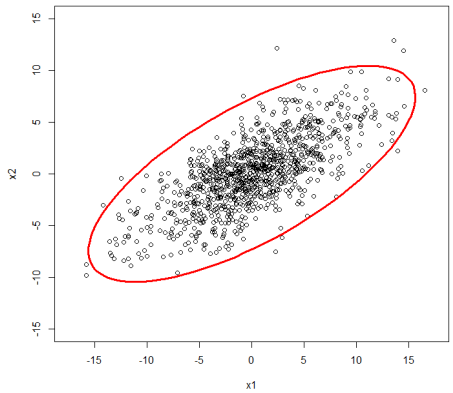
A Shewhart control chart



Univariate charts

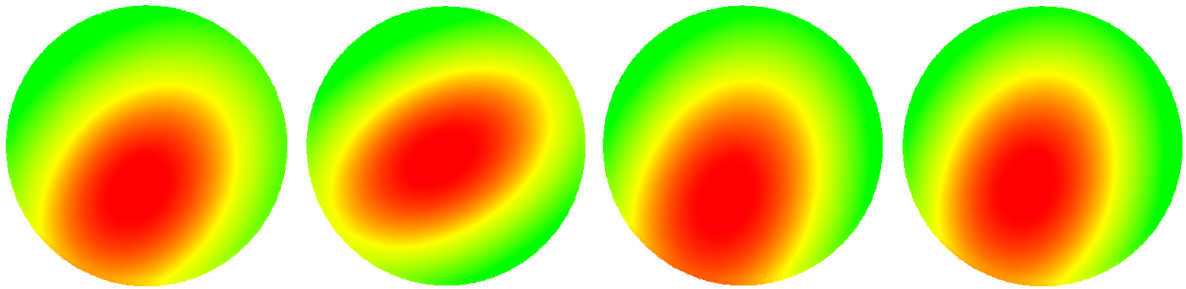


Principal components



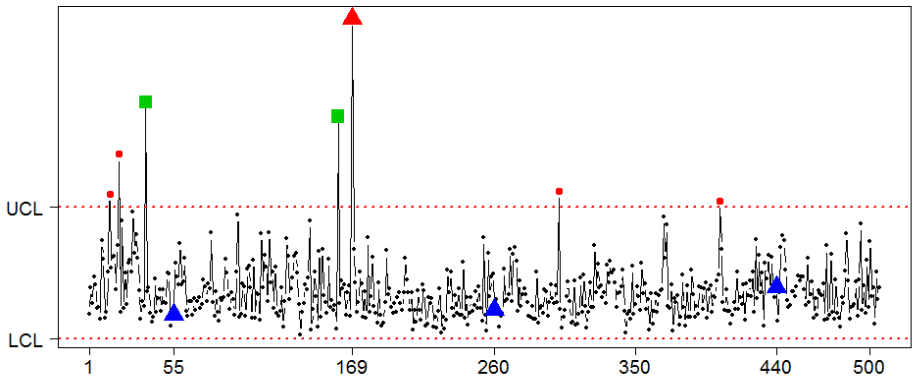
Hotelling's T^2

Example 2. To detect anomalies in wafers



Wafer 55 Wafer 169 Wafer 260 Wafer 440

Profiles of wafers marked with triangles

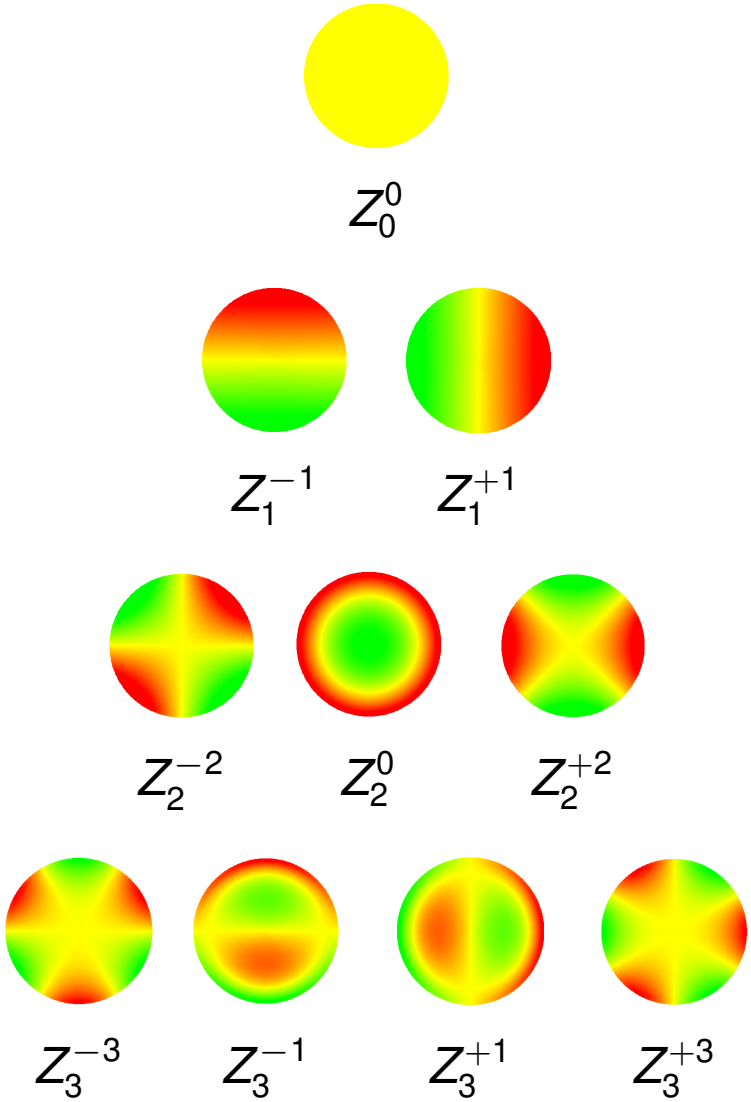


Control chart for 506 wafers

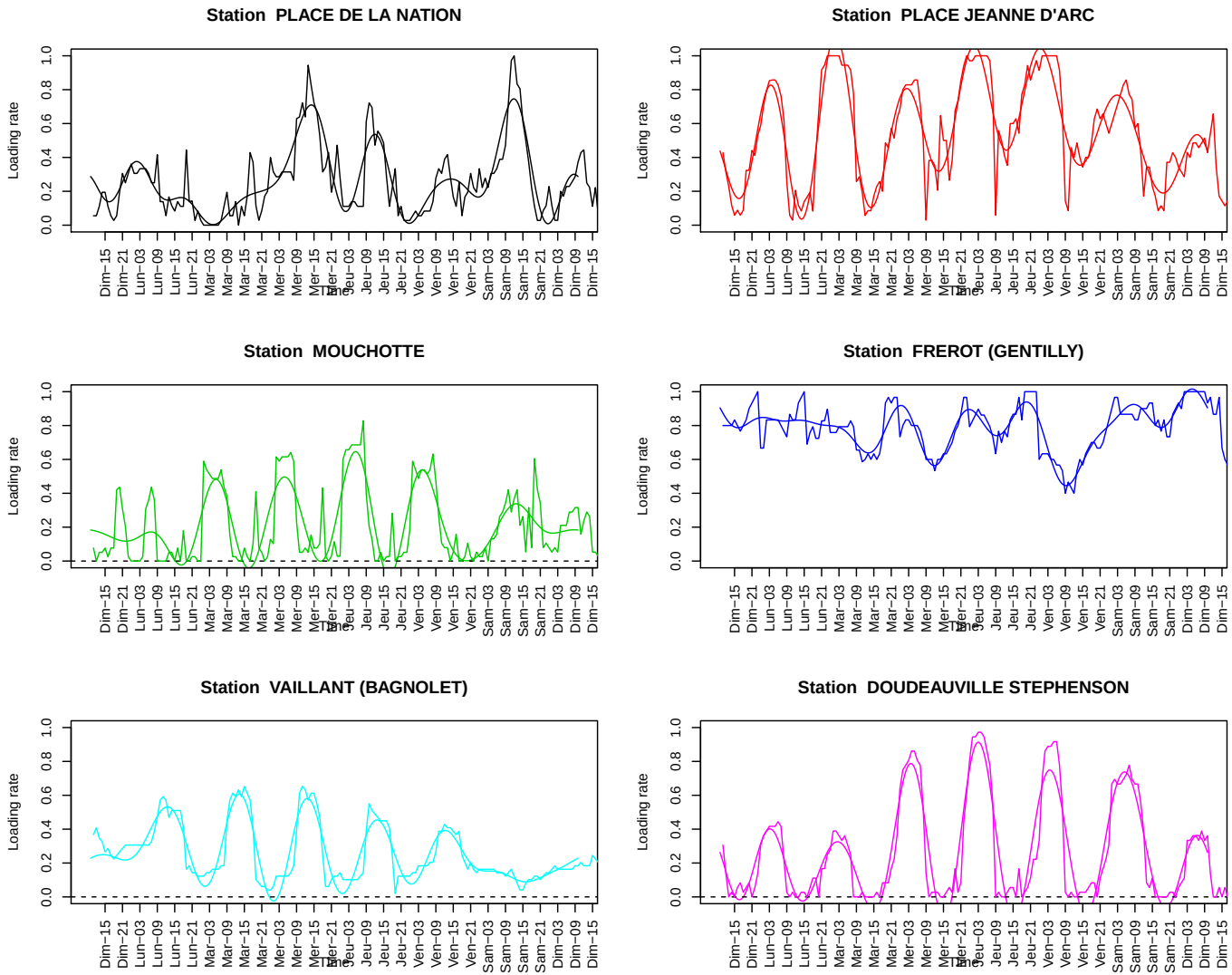
Example 2. To detect anomalies in wafers

Control charts have been applied to the **coefficients in a functional basis**,

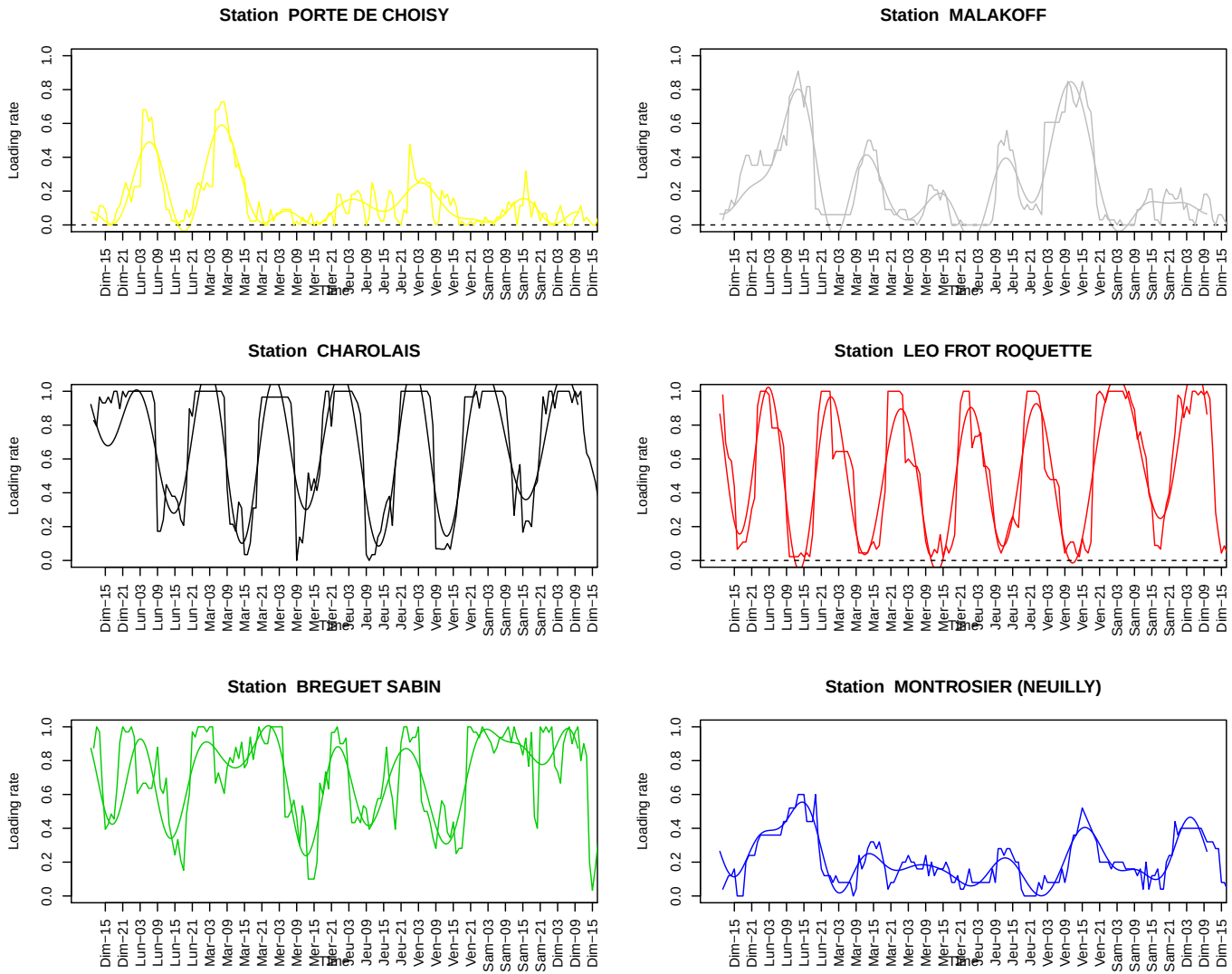
here: Zernike polynomials



Example 3. Data exploration: To detect patterns in velib



Example 3. Data exploration: To detect patterns in velib



Example 3. Data exploration: To detect patterns in velib

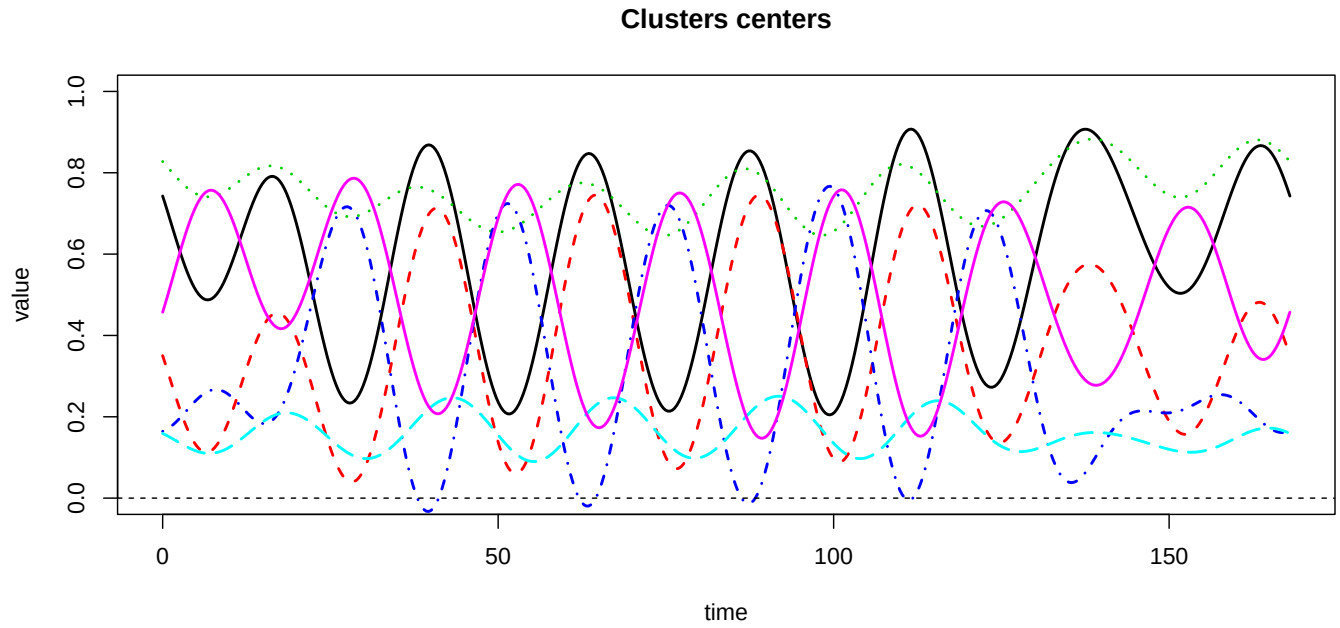


Figure: Clusters found by k-means, applied on the coefficients of Fourier series

Example 3. Data exploration: To detect patterns in velib

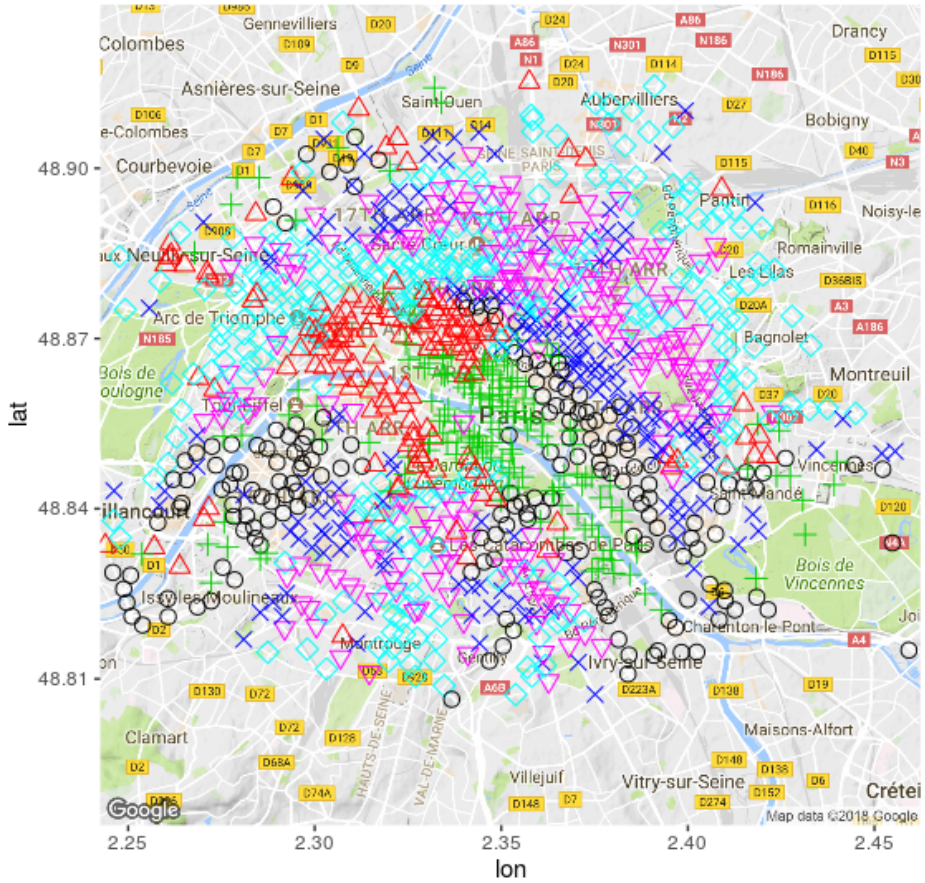


Figure: Visualization of velib stations, gathered by cluster