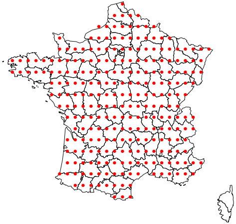
Graph Functional Methods for Climate Partitioning

Mathilde Mougeot
-
with D. Picard, V. Lefieux*, M. Marchand*

Université Paris Diderot, France
*Réseau Transport Electrique (RTE)

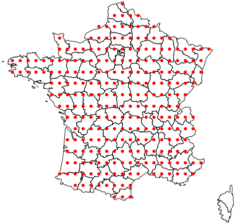
Buenos Aires, 2015

Arpège , French Meteorological data

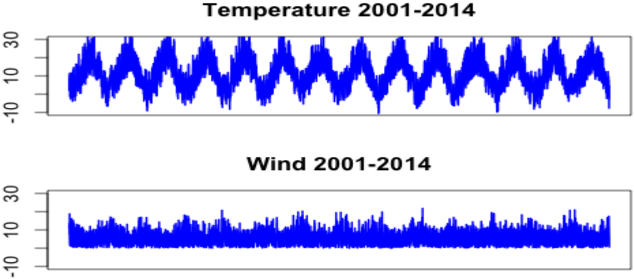


At $n = 259$ locations, we observe
- Temperature and Wind data
- for 14 years
- with an hourly sample rate
- $d = 122\,712$ points for raw data
- X matrix of data (n x d) $n < d$

Arpège , French Meteorological data



At $n = 259$ locations, we observe
- Temperature and Wind data
- for 14 years
- with an hourly sample rate
- $d = 122\,712$ points for raw data
- X matrix of data (n x d) $n < d$



Objective and Questions :

RTE requirement

- Segmentation of the French country using meteorological data
- Temperature and/or Wind
- To study the Between Year variability, we focus on
    - 14 x one year of data (n=259 x d=8760-daily) (vs 14 years of data)

Objective and Questions :

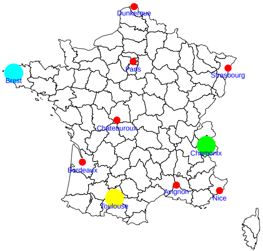
RTE requirement

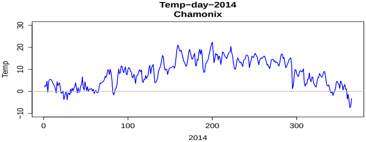
- Segmentation of the French country using meteorological data
- Temperature and/or Wind
- To study the Between Year variability, we focus on
  - 14 x one year of data (n=259 x d=8760-daily) (vs 14 years of data)

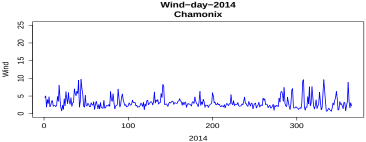
Methodological & Statistical Questions :

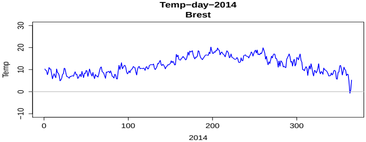
- High dimensional data $n = 259$, $d >> n$
- to avoid the curse of dimensionality
  Features extraction, Smoothing, and/or temporal aggregation
- Clustering algorithms :
  - ? Hierarchical clustering, Kmeans, Spectral clustering among others
  - ? number of clusters
- How to aggregate the clustering results between years ?

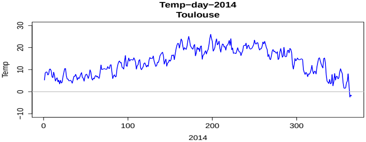
Wind and Temperature data spots for 2014

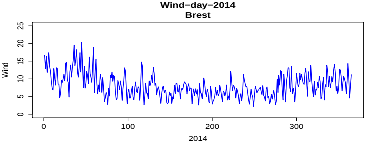


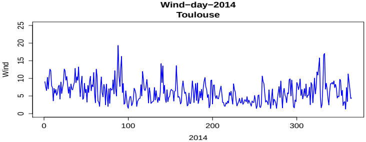












Features Extraction

From Temporal time series to feature

Non Parametric Regression

Principal Component Analysis

to avoid the curse of dimensionality with clustering

Feature Extraction based on Non Parametric Regression

Original data are function of time observed at regular instances.
For each time series at location i of size d, we observe $(X_t^i, t/d)$ where

$$X_t^i = f^i(t/d) + \epsilon_t^i,$$

f^i is unknow, $\epsilon^i \sim \mathcal{N}(0, \sigma^2)$,
$t = 1, \ldots, d$.
Non parametric estimation of f^i :

$f^i = \sum_{\ell=1}^{p} \beta_\ell^i g_\ell + h^i$

with $\mathcal{D} = \{g_1, \ldots, g_p\}$ dictionary of functions.
G is the (d, p) design matrix

Mougeot et al. JRSSB, 2012

Feature Extraction based on Non Parametric Regression

Original data are function of time observed at regular instances.
For each time series at location i of size d, we observe $(X^i_t, t/d)$ where

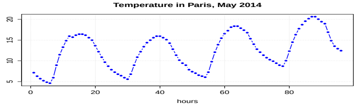
$$X^i_t = f^i(t/d) + \epsilon^i_t,$$

f^i is unknow, $\epsilon^i \sim \mathcal{N}(0, \sigma^2)$,
$t = 1, \ldots, d$.
Non parametric estimation of f^i :

$$f^i = \sum_{\ell=1}^{p} \beta^i_\ell g_\ell + h^i$$

with $\mathcal{D} = \{g_1, \ldots, g_p\}$ dictionary of functions.
G is the (d, p) design matrix

Mougeot et al. JRSSB, 2012



OLS : $\hat{\beta}^i = \arg\min_\beta ||X^i - G\beta^i||^2$

Thresholding and Sparsity :

We note $\hat{X}^i_{j_0} = \sum_{j=1}^{j_0} \hat{\beta}^i_{(j)} g_j$
with $|\hat{\beta}^i_{(1)}| \geq \ldots \geq |\hat{\beta}^i_{(n)}|$, and

$$\frac{||\hat{X}^i_{(j_0)}||^2}{||X^i||^2} \geq T_{NP} (= 0.95).$$

Feature (Sparse) matrix :

$$\boxed{Z = (Z_{i,j}) = (\hat{\beta}_{i,j})}$$

Feature Extraction using Principal Component Analysis

Projection of the observations using a data driven orthonormal basis

X centered data matrix (n, d)
$n = 259$, $d >> n$ large

The Feature matrix (n, p) is computed
by projection, $p << d$:

$$\boxed{Z = XU_p}$$

U_p is the matrix defined by the first eigenvectors of the S.
S is the Variance-Covariance matrix.

Feature Extraction using Principal Component Analysis

Projection of the observations using a data driven orthonormal basis

X centered data matrix (n, d)
$n = 259$, $d >> n$ large

The Feature matrix (n, p) is computed by projection, $p << d$:

$$\boxed{Z = XU_p}$$

U_p is the matrix defined by the first eigenvectors of the S.
S is the Variance-Covariance matrix.

$S = \frac{1}{n} X^T X$

$S = U_d \Sigma_d U_d^T$ (SVD)
with $\Sigma_d = diag(\lambda_1, ...\lambda_d)$
$\lambda_1 \geq \ldots, \geq \lambda_d$ eigenvalues

This transformation maximizes the variance on the different axes.

How to chose p ? :

p such that $\frac{\lambda_1 + \ldots + \lambda_p}{\Sigma_j \lambda_j} = T_{pca}$ (0.95)

$\rightarrow$ Global linear method involving all the $n = 259$ spots to compute U_p
$\rightarrow$ Is U_p similar between years ?

Number of extracted features

Average nb. of Features for Temperature and Wind over 14 years

Temperature	day (d=365)	week (52)	month (12)
PCA 95%	12.3 (1.1)	5.65 (0.23)	3 (0.1)
PCA "leave one out"	243 (2.4)	21 (0.8)	4.68 (0.4)
NP Reg. Fourier d.	127 (3.4)	19.5 (0.8)	4.45 (0.3)
NP Reg. Haar d.	138 (3.9)	19.7 (0.9)	5.78 (0.3)

Wind	day (365)	week (52)	month (12)
PCA 90%	14.1 (2.5)	6.36 (1.2)	2.57 (0.5)
PCA leave on out	258 (0)	32.4 (02.17)	5.17 (0.9)
NP Reg. Fourier d.	233 (6.8)	31.7 (1.9)	5.14 (0.77)

→ Sparse representation of PCA for daily Temperature

→ PCA projection matrix can not be learned

→ Similar nb. of features for PCA and generic Fourier dico for monthly data

Clustering algorithms

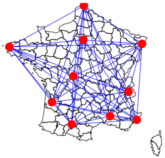
Hierarchical clustering
Kmeans
Spectral clustering

–

Aggregation of clustering instances

Spectral clustering

Full connected graph with n nodes.



Weight between two nodes (Z_i, Z_j) :
$w_{i,j} = e^{\frac{-||Z_i - Z_j||_2^2}{2\mu^2}}$,
μ heat parameter

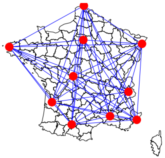
Normalized Graph Laplacian :

$$\boxed{L = I - D^{-1/2} W D^{-1/2}}$$

$L \in \mathbb{R}^{N \times N}$,
W adjacency matrix, $D_{i,i} = \sum_j w_{i,j}$.

Spectral clustering

Full connected graph with n nodes.



Weight between two nodes (Z_i, Z_j) :

$w_{i,j} = e^{\frac{-||Z_i - Z_j||_2^2}{2\mu^2}}$,

μ heat parameter

Normalized Graph Laplacian :

$$\boxed{L = I - D^{-1/2} W D^{-1/2}}$$

$L \in \mathbb{R}^{N \times N}$,

W adjacency matrix, $D_{i,i} = \sum_j w_{i,j}$.

Ng et al. Algorithm (2002) :

Input : Fix k nb . clusters

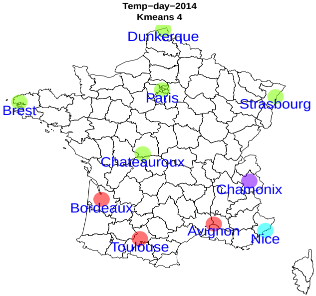
1. Compute the first k eigenvectors $u_1, \dots u_k$ of L corresponding to the "k" smallest eigenvalues,
2. let $U \in \mathbb{R}^{n \times k}$ be the matrix of column vectors $u_1, \dots, u_k$.
3. Form the matrix $T \in \mathbb{R}^{n \times k}$ $t_{i,j} = u_{i,j} / (\sqrt{\sum_k u_{ik}^2})$. Let $y_i \in R^k$ i^{th} row of T.
4. Cluster $\{y_i\}$, $1 \le i \le n$ with the k-means into clusters $C_1, \dots, C_k$

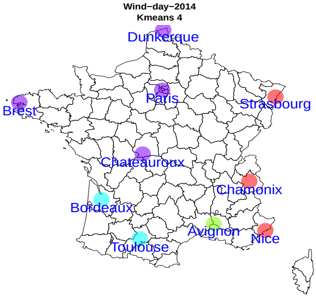
Output : Clusters $A_1, \dots, A_k$ with $A_i = \{y_i \in C_i\}$

Kmeans Clustering

Choose k the number of clusters

1. INPUT Pre specify k centroids $\bar{Z}_1 \dots \bar{Z}_k$ (k points at random)
2. Reassign each item to its nearest cluster centroid
3. Compute the Squared Euclidien Distance $ESS = \sum_{k=1}^{K} \sum_{c(i)=k} ||Z_i - \bar{Z}_k||^2$
4. Update the cluster centroids after each assignment.
5. REPEAT 2,3,4 with UNTIL no further assignment of items takes place. (or a given nb. of runs)





How to choose the number of clusters ?

Many methods already in the literature : Calinsky et al. 1974, Gap Statistic Friedman et al. 2000, ... Most of them based on :

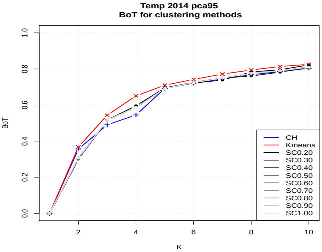
Variance Decomposition : $T = W_k + B_k$

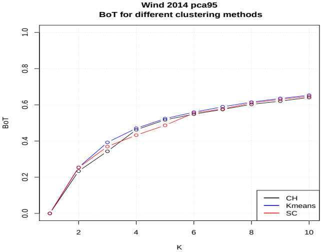
Total	$T = \frac{1}{n}\sum_i \vert\vert X_i - \bar{X}\vert\vert^2$
Between	$B_k = \frac{1}{n}\sum_k n_k \vert\vert \bar{X}_k - \bar{X}\vert\vert^2$
Within	$W_k = \frac{1}{n}\sum_k \sum_{i_k}^{n_k} \vert\vert X_k(i_k) - \bar{X}_k\vert\vert^2$

Quantification/ modeling indicator ratio :

$$\boxed{\rho_k = \frac{B_k}{T} \in [0, 1]}$$

k_0 the number of cluster is chosen such that :
with $\Delta_k = \rho_{k+1} - \rho k$
$k_0 = \arg\min_k \Delta_k < 5\%$





Daily Temperature 2014

Application to segmentation

&

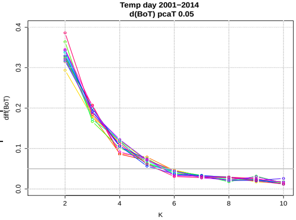
Numerical Results

Stability of the number of clusters

over 14 years, for different temporal aggregation levels

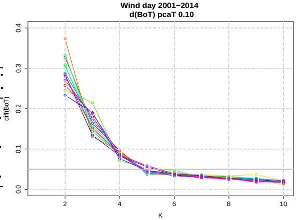
Data : 14 x one year of data,
Kmeans algorithm with $\rho_k < 5\%$ criteria

Temperature :



	day (365)	week (52)	month (12)
PCA 95%	5 (0)	4.9 (0.2)	4.7 (0.4)
NP Reg. Trigo	5 (0)	4.8 (0.4)	4.7 (0.4)
NP Reg. Haar	5 (0)	4.8 (0.4)	4.7 (0.4)

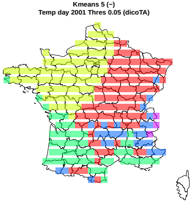
Wind :

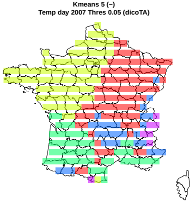


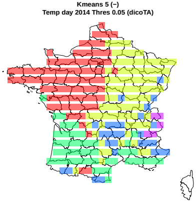
	day (365)	week (52)	month (
Pca 90%	4.15 (0.3)	4.23 (0.4)	4.31 (0.
NP Reg. Trigo	4.15 (0.3)	4 (0)	4.08 (0.
NP Reg. Haar	4.23 (0.4)	4.31 (0.4)	4.15 (0.

Segmentation for 2001, 2007, 2014 daily data, $n = 259$

Temperature

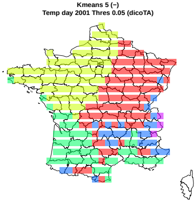


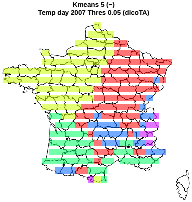


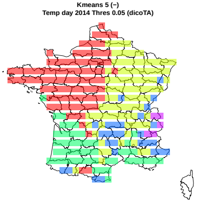


Segmentation for 2001, 2007, 2014 daily data, $n = 259$

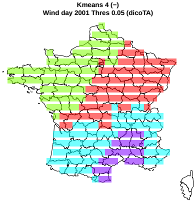
Temperature

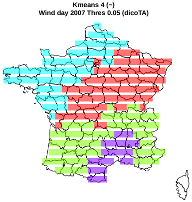


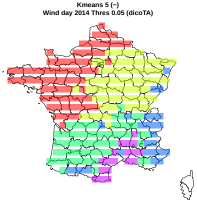




Wind

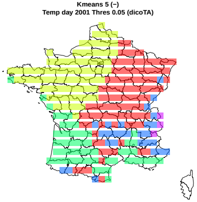


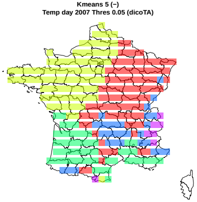


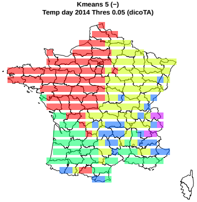


Segmentation for 2001, 2007, 2014 daily data, $n = 259$

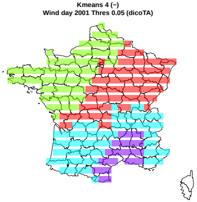
Temperature

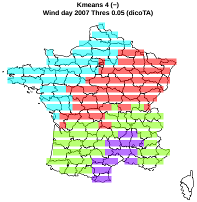


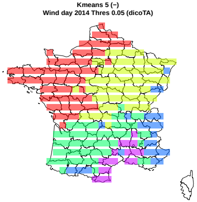




Wind







Next step : aggregation of clustering instances ?

Aggregation of clustering instances

"Cluster Ensemble" approach :

For $y \in \mathcal{Y} = \{2001, \ldots, 2014\}$
1 : Feature extraction Z^y for year y
2 : Clustering using $\mathcal{M}$, $\mathcal{M} \in$ {HC, Kmeans, SC}
3 : Construct a co-cluster indicator matrix A^y

$$A^y_{i,j} = \begin{cases} 1 \text{ if } (Z^y_i, Z^y_j) \text{ in the same cluster} \\ 0 \text{ otherwise} \end{cases}$$

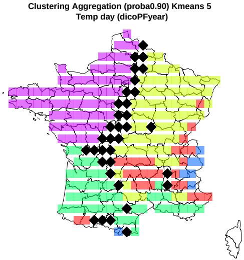
EndFor

Averaged co cluster indicator matrices :
$A \leftarrow \frac{1}{\sharp\mathcal{Y}} \sum_y A^y$

$$A_{i,j} = \begin{cases} 1 \ (Z_i,\ Z_j) \ \textbf{always} \text{ in the same cluster} \\ 0 \ (Z_i,\ Z_j) \ \textbf{never} \text{ in the same cluster} \end{cases}$$

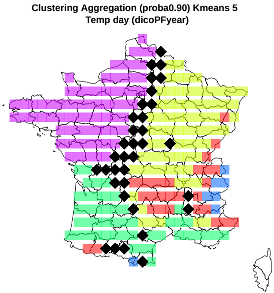
$(1 - A)$ is an affinity matrix.

Spectral Clustering :
$w_{i,j} = e^{\frac{-(1-A_{i,j})}{2\mu^2}}$,

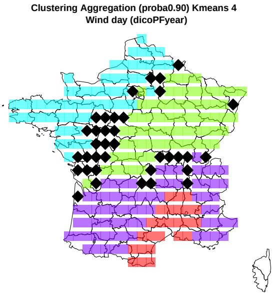


"Behavior index" :
$\pi_i = \frac{1}{n} \sum_j 1_{\{\epsilon < A_{i,j} < (1-\epsilon)\}}$,
$\epsilon = 0.10$

Aggregation of clustering instances



Temperature



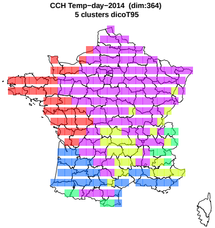
Wind

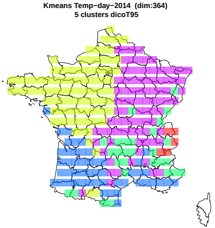
using 14 years of daily data

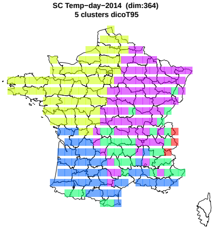
Impact of clustering methods on segmentation

What choice between Hierarchical clustering, Kmeans, Spectral Clustering ?

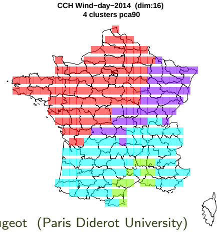
Temperature

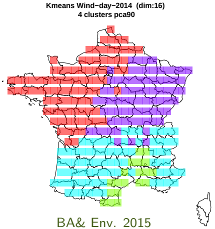


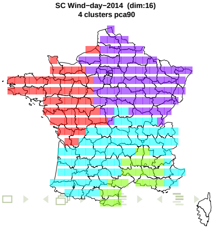




Wind







Impact of Feature extraction on Clustering results

Quantification on the impact of Feature extraction using a Cluster Ensemble approach :

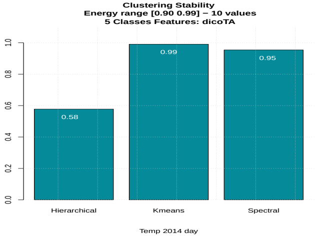
For $T \in \{0.90, \dots 0.99\}$
1 : Compute Feature Z , with T
2 : Features Clustering using $\mathcal{M}$
3 : Construct a co-cluster indicator matrix A^T

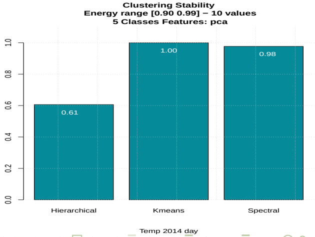
$$A^t_{i,j} = \begin{cases} 1 \text{ if } Z_i \text{ and } Z_j \text{ are in the same cluster} \\ 0 \text{ otherwise} \end{cases}$$

EndFor
$A \leftarrow \frac{1}{\sharp T} \sum_{t \in T} A^t$ (average indicator matrices)

$$Ratio = \frac{1}{(n(n-1)/2)} \sum_{i<j} 1_{0<A_{i,j}<1}$$

$\mathcal{M} \in \{\text{HC, Kmeans, SC}\}$





$\rightarrow$ Hierarchical clustering results are very sensitive to Feature extraction parameter.

Conclusion

- Clustering is, at the same time,
  - an easy task (just apply a clustering method and see what happen !),
  - a very hard one (no objective functions as MSE)

Conclusion

- Clustering is, at the same time,
  - an easy task (just apply a clustering method and see what happen !),
  - a very hard one (no objective functions as MSE)
- To be very careful with
  - potential high dimensional data as the ℓ_2 norm may not be meaningfull
  - the robustness on the results provided by the pre treatments (smoothing)

Conclusion

- Clustering is, at the same time,
  - an easy task (just apply a clustering method and see what happen !),
  - a very hard one (no objective functions as MSE)
- To be very careful with
  - potential high dimensional data as the ℓ_2 norm may not be meaningfull
  - the robustness on the results provided by the pre treatments (smoothing)
- We propose a methodology based on :
  - Feature extraction (PCA, Non Parametric Regression)
  - Clustering a set of data (split the initial time series into 14 one year intervals)
  - Aggregate the clustering with alike "Ensemble method" and Spectral clustering

Conclusion

- Clustering is, at the same time,
  - an easy task (just apply a clustering method and see what happen !),
  - a very hard one (no objective functions as MSE)
- To be very careful with
  - potential high dimensional data as the ℓ_2 norm may not be meaningfull
  - the robustness on the results provided by the pre treatments (smoothing)
- We propose a methodology based on :
  - Feature extraction (PCA, Non Parametric Regression)
  - Clustering a set of data (split the initial time series into 14 one year intervals)
  - Aggregate the clustering with alike "Ensemble method" and Spectral clustering
- Work still in progress
  - to quantify the benefits of NP regression modeling compared to PCA
  - To cluster at the same time Temperature and Wind data

Thank you for your attention