Non supervised classification of individual electricity curves

Jairo Cugliari

17 février 2014

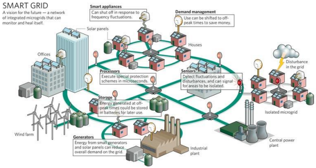
Joint work with Benjamin Auder (LMO, Université Paris-Sud)

Outline

Sommaire

Industrial motivation

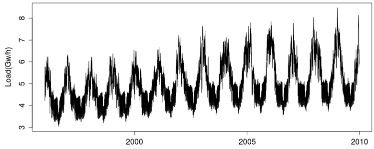
- Smartgrid & Smart meters : time real information
- Lot of data of different nature
- Many problems : transfer protocol, security, privacy, ...
- The French touch : 35M Linky smartmeter



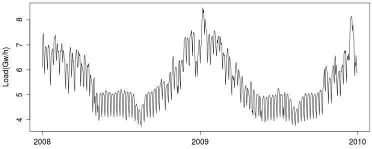
What can we do with all these data ?

Electricity demand data

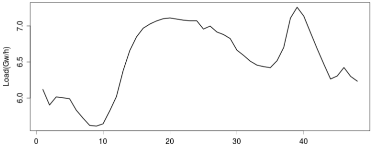
Some salient features



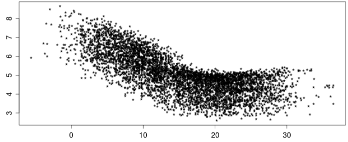
(a) Long term trand



(b) Weekly cycle



(c) Daily load curve



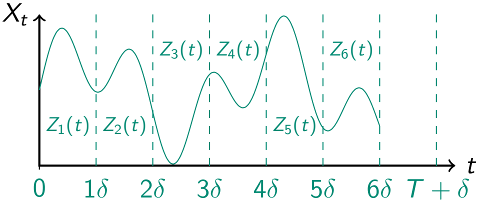
(d) Electricity load vs. temperature

FD as slices of a continuous process

[Bosq, (1990)]

The prediction problem

- Suppose one observes a square integrable continuous-time stochastic process $X = (X(t), t \in \mathbb{R})$ over the interval $[0, T]$, $T > 0$;
- We want to predict X all over the segment $[T, T+\delta], \delta > 0$
- Divide the interval into n subintervals of equal size δ.
- Consider the functional-valued discrete time stochastic process $Z = (Z_k, k \in \mathbb{N})$, where $\mathbb{N} = \{1, 2, \dots\}$, defined by

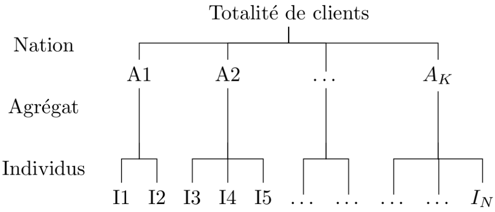


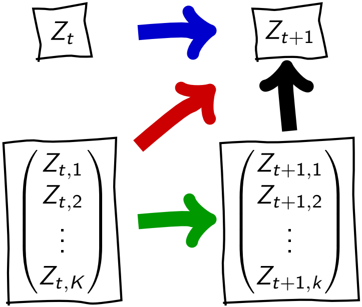
$$Z_k(t) = X(t + (k-1)\delta)$$

$$k \in \mathbb{N} \quad \forall t \in [0, \delta)$$

If X contents a $\delta-$seasonal component, Z is particularly fruitful.

Long term objective





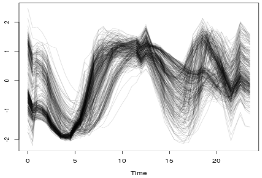
- Groups can express tariffs, geographical dispersion, client class ...
- **IDEA** : Use a clustering algorithm to learn groups of customer structure
- **Aim** : Set up a classical clustering algorithm to run in parallel

Sommaire

Aim

- Segmentation of X may not suffices to render reasonable the stationary hypothesis.
- If a grouping effect exists, we may considered stationary within each group.
- Conditionally on the grouping, functional time series prediction methods can be applied.
- We propose a clustering procedure that discover the groups from a bunch of curves.

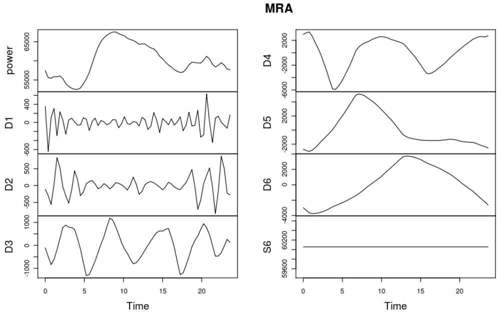
We use wavelet transforms to take into account the fact that curves may present non stationary patters.



Two strategies to cluster functional time series :

1. Feature extraction (summary measures of the curves).
2. Direct similarity between curves.

Wavelets to cope with FD



- domain-transform technique for hierarchical decomposing finite energy signals
- description in terms of a broad trend (approximation part), plus a set of localized changes kept in the details parts.

Discrete Wavelet Transform

If $z \in L_2([0,1])$ we can write it as

$$z(t) = \sum_{k=0}^{2^{j_0}-1} c_{j_0,k}\phi_{j_0,k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} d_{j,k}\psi_{j,k}(t),$$

where $c_{j,k} =< g, \phi_{j,k} >$, $d_{j,k} =< g, \varphi_{j,k} >$ are the scale coefficients and wavelet coefficients respectively, and the functions ϕ et φ are associated to a orthogonal MRA of $L_2([0,1])$.

Energy decomposition of the DWT

- Energy conservation of the signal

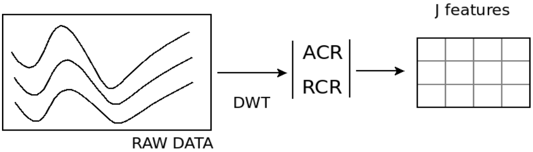
$$\|z\|_H^2 \approx \|\widetilde{z}_J\|_2^2 = c_{0,0}^2 + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} d_{j,k}^2 = c_{0,0}^2 + \sum_{j=0}^{J-1} \|\mathbf{d}_j\|_2^2.$$

- For each $j = 0, 1, \ldots, J-1$, we compute the absolute and relative contribution representations by

$$\underbrace{\text{cont}_j = ||\mathbf{d_j}||^2}_{\boxed{\text{AC}}} \qquad \text{and} \qquad \underbrace{\text{rel}_j = \frac{||\mathbf{d_j}||^2}{\sum_j ||\mathbf{d_j}||^2}}_{\boxed{\text{RC}}}.$$

- They quantify the relative importance of the scales to the global dynamic.
- RC normalizes the energy of each signal to 1.

Schema of procedure



0. Data preprocessing. Approximate sample paths of $z_1(t), \ldots, z_n(t)$
1. Feature extraction. Compute either of the energetic components using absolute contribution (AC) or relative contribution (RC).
2. Feature selection. Screen irrelevant variables. [Steinley & Brusco ('06)]
3. Determine the number of clusters. Detecting significant jumps in the transformed distortion curve. [Sugar & James ('03)]
4. Clustering. Obtain the K clusters using PAM algorithm.

1. Antoniadis, X. Brossat, J. Cugliari et J.-M. Poggi (2013), Clustering Functional Data Using Wavelets, *IJWMIP*, 11(1), 35–64

Sommaire

Partitioning Around Medoids (PAM) [Kaufman et Rousseeuw (1987)]

- Partition the n points R^d-scatter into K clusters
- Optimization problem :

$$D(x) = \min_{m_1,\dots,m_k \in \mathbb{R}^d} \sum_{i=1}^{n} \min_{j=1,\dots,k} \|x_i - m_j\| ,$$

  with $x = (x_1, \dots, x_n)$, $\| . \|$ can be any norm. Here we choose to use the euclidean norm.
- Robust version of k-means
- Computational burden : medians instead of means
- Several heuristics allow to reduce the computation time.

Parallelization with MPI

- Easy to use library routines allowing to write algorithms in parallel
- Available on several languages
- We use the master-slave mode

The outline of code :

1. The master process splits the problem in tasks over the data set and sends it to the workers ;
2. Each worker reduces the functional nature of the data using the DWT, applies the clustering and returns the centers ;
3. The master recuperates and clusters the centers into K meta centers.

The source code is open and will be available to download from github.

---

1. B. Auder & J. Cugliari. Parallélisation de l'algorithme des k-médoïdes. Application au clustering de courbes. (2014, submitted)

Sommaire

Application I : Starlight curves

- Data from UCR Time Series Classification/Clustering
- 1000 curves learning set + 8236 validation set ($d = 1024$)

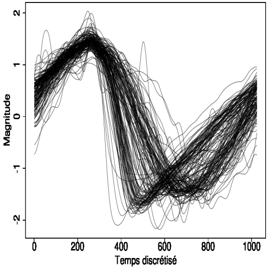


FIGURE: Groupe 1

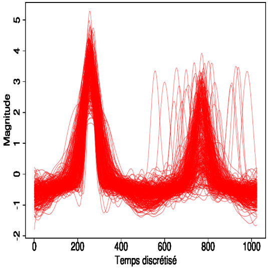


FIGURE: Groupe 2

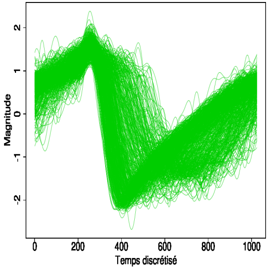


FIGURE: Groupe 3

		Adequacy	
	Distortion	Internal	External
Training (sequential)	1.31e4	0.79	0.77
Training (parallel)	1.40e4	0.79	0.68
Test (sequential)	1.09e5	0.78	0.76
Test (parallel)	1.15e5	0.78	0.69

Application II : EDF data

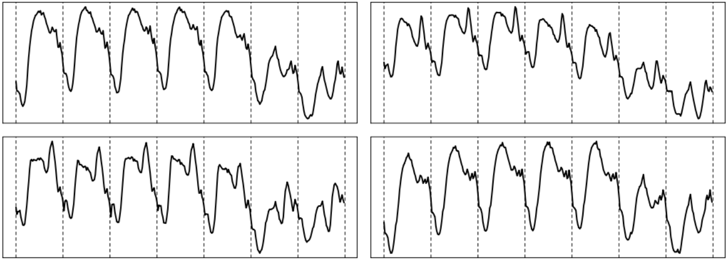


FIGURE: French electricity power demand on autumn (top left), winter (bottom left), spring (top right) and summer (bottom right).

Feature extraction :

- The significant scales for revealing the cluster structure are independent of the possible number of clusters.
- Significant scales are associated to mid-frequencies.
- The retained scales parametrize the represented cycles of 1.5, 3 and 6 hours (AC).

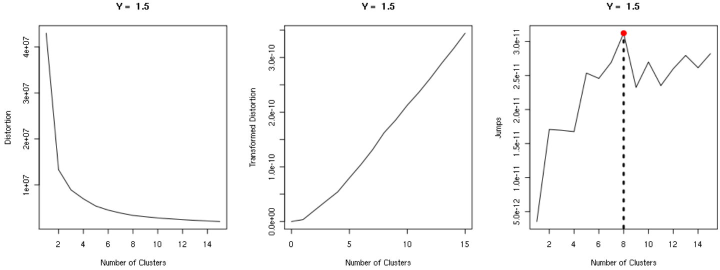
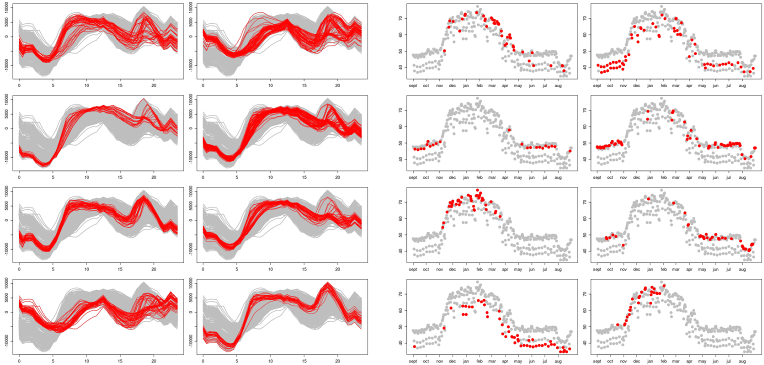


FIGURE: Number of clusters by feature extraction of the AC (top). From left to right : distortion curve, transformed distortion curve and first difference on the transformed distortion curve.



(a) Cluster (b) Calendar

FIGURE: Curves membership of the clustering using AC based dissimilarity (a) and the corresponding calendar positioning (b).

Application III : Electricity Smart Meter CBT (ISSDA)

- 4621 Irish households smart meter data
- About 25K discretization points
- We test with $K = 3$ or 5 classes
- We compare sequential and parallel versions

	Distortion	Internal adequacy
3 clusters sequential	1.90e7	0.90
3 clusters parallel	2.15e7	0.90
5 clusters sequential	1.61e7	0.89
5 clusters parallel	1.84e7	0.89

1. *Irish Social Science Data Archive*, http://www.ucd.ie/issda/data/

Sommaire

Conclusion

- Identification of customers groups from smartmeter data
- Wavelets allow to capture the functional nature of the data
- Clustering algorithm upscale envisaged for millions of curves
- *Divide-and-Conquer* approach thanks to MPI library

Further work

- Go back to the prediction task
- Apply the algorithm over many hundreds of processors
- Connect the clustering method with a prediction model

Bibliographie

- A. Antoniadis, X. Brossat, J. Cugliari et J.-M. Poggi (2013), Clustering Functional Data Using Wavelets, *IJWMIP*, 11(1), 35–64

- R. Bekkerman, M. Bilenko et J. Langford - éditeurs (2011), Scaling up Machine Learning : Parallel and Distributed Approaches, *Cambridge University Press*

- P. Berkhin (2006), A Survey of Clustering Data Mining Techniques, *Grouping Multidimensional Data, éditeurs : J. Kogan, C. Nicholas, M. Teboulle*.

- J. Dean et S. Ghemawat (2004), MapReduce : Simplified Data Processing on Large Clusters, *Sixth Symposium on Operating System Design and Implementation*.

- G. De Francisci Morales et A. Bifet (2013), G. De Francisci Morales SAMOA : A Platform for Mining Big Data Streams Keynote Talk at RAMSS '13 : 2nd International Workshop on Real-Time Analysis and Mining of Social Streams WWW, Rio De Janeiro

- L. Kaufman et P.J. Rousseeuw (1987), Clustering by means of Medoids, *Statistical Data Analysis Based on the L_1-Norm and Related Methods, éditeur : Y. Dodge*.