Disaggregated Electricity Forecasting using Wavelet-based Clustering of Individual Consumers

Jean-Michel POGGI

UNIVERSITÉ PARIS SUD

UNIVERSITÉ PARIS DESCARTES

Modelling Smart Grids 2018, Prague, September 20-22, 2018

This talk is mainly related to joint works with A. Antoniadis (Univ. Grenoble, France & Univ. Cape Town, South Africa), X. Brossat (EDF R&D, France), J. Cugliari (Univ. Lyon 2, France) and Y. Goude (EDF R&D, France and Univ. Paris-Sud, Orsay, France) and a long scientific collaboration with EDF (Electricité de France)

Industrial motivation

Short-term electricity demand forecast

- Time series analysis : SARIMA(X), Kalman filter [Dordonnat *et al.* (2009)]
- Machine learning [Devaine *et al.* (2010)]
- Similarity search based methods [Poggi (1994), Antoniadis *et al.* (2006)]
- Regression : EDF modelisation scheme [Bruhns *et al.* (2005)], GAM [Pierrot and Goude (2011)], Bayesian approach [Launay, Philippe and Lamarche (2012)]

New challenges in electricity demand forecast

- Market liberalization : may produce variations on clients' perimeter that risk to induce nonstationarities on the signal
- Development of smart grids and smart meters : various aggregates are simultaneously of interest.

Needs to build models relying more deeply to the multiscale nature of the data, both in time and space

Short bibliography about bottom-up forecasting

- *Bottom-up forecasting predicts the total consumption of a set of customers using individual metered data.*
- *Combines clustering methods to build clusters of customers, forecasting models in each cluster and then aggregating them.*
- Clustering can be quite independent of the forecasting model (Irish dataset) [Alzate, Sinn, Proc. of the Int. Conf. on Data Mining, 2013.]
- On the same dataset, a longitudinal clustering and a functional forecasting model similar to our KWF for forecasting individual load curves in [Chaouch, M., IEEE Trans. Smart Grid 2014.]
- A clustering method supervised by a forecasting accuracy, applied to a French industrial subset obtaining a substantial gain but computationally intensive, in [Misiti, Misiti, Oppenheim, Poggi, Rev. Stat. J. 2010, 8, 105–124.]
- A k-means based on electrical features and, in each cluster, deep learning is used for forecasting. A minimum gain of 11% in forecast accuracy on the Irish data set and smart meter data from New-York [Quilumba, Lee, Huang, Wang, Szabados, IEEE Trans. Smart Grid 2015]
- Ensemble of forecasts from a hierarchical clustering on individual average weekly profiles, coupled with a deep learning model for forecasting in each cluster, and at the end combined using linear regression [Wang, Chen, Hong, Kang, IEEE Trans. Smart Grid 2018.]

1. Antoniadis, Brossat, Cugliari, Poggi, J. Soc. Française Stat. 2012, 153, 52–78.

Clients hierarchical structure and prediction

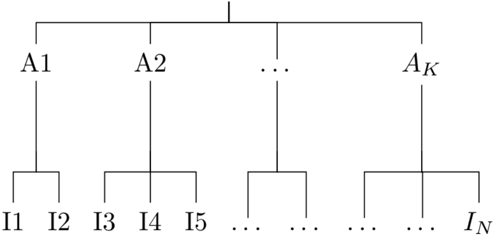
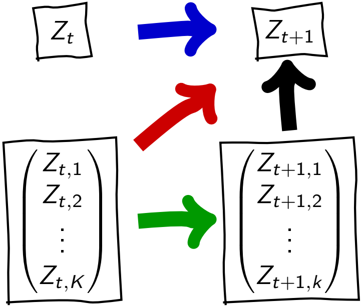


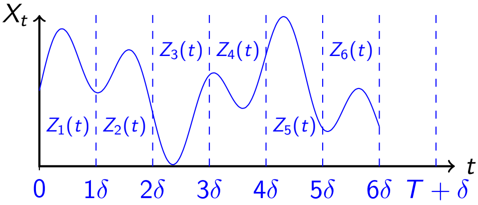
FIGURE : Hierarchical structure of N individual clients among K groups.



- Z_t : aggregate demand at t
  $Z_{t,k}$: demand of group k at moment t
- Groups can express tariffs, geographical dispersion, client class ...
- Profiling *vs* Prediction
- We follow Misiti *et al.* (2010) to construct clusters of customers to better predict the global aggregate

Functional data as slices of a continuous process

- Observe 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 $\mathbb{N} = \{1, 2, \ldots\}$, defined by

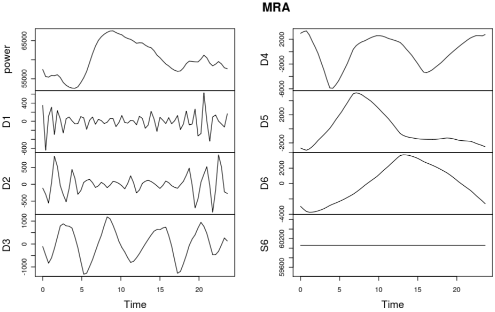


$$Z_k(t) = X(t + (k - 1)\delta)$$

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

- If X exhibits a $\delta-$seasonal component, Z is particularly fruitful, transferring complexity to intraday features

Wavelets to cope with Functional Data



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

Discrete Wavelet Transform (DWT)

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 through DWT

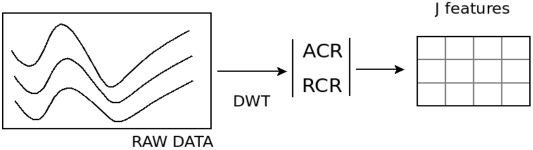
- Energy conservation of the signal

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

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

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

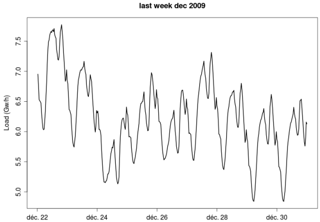
A first clustering procedure

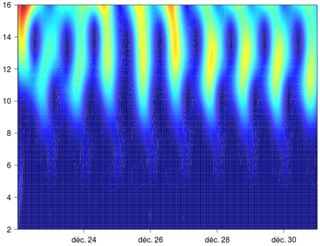


0. Data preprocessing. Approximate sample paths of $z_1(t), \ldots, z_n(t)$
1. Feature extraction. Compute either of the energetic components using absolute contribution (AC) or relative contribution (RC).
2. Feature selection. Screen irrelevant variables. [Steinley & Brusco ('06)]
3. Determine the number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

A fully time-scale function-based dissimilarity

- Distance based on wavelet-correlation between two time series
- Wavelet coherence provides a generalization of local Fourier cross-spectrum
- Can be used to measure relationship between two functions
- The strength of the relation is hierarchically decomposed across scales without losing of time location
- A major drawback is that it needs more computation time and storage (complex values)





Wavelet coherence (1/2)

Continuous Wavelet Transform (CWT)

Starting with a mother wavelet ψ consider $\psi_{a,\tau} = a^{-1/2}\psi\left(\frac{t-\tau}{a}\right)$.
The CWT of a function $z \in L^2(\mathbb{R})$ is,

$$W_z(a,\tau) = \int_{-\infty}^{\infty} z(t)\psi^*_{a,\tau}(t)dt$$

As for Fourier transform, a spectral approach is possible.

$$S_z(a,\tau) = |W_z(a,\tau)|^2 \qquad \text{(wavelet spectrum)}$$
$$\mathcal{W}_{z,x}(a,\tau) = W_z(a,\tau)W^*_x(a,\tau) \qquad \text{(cross-wavelet transform)}$$

Wavelet coherence (2/2)

$$R^2_{z,x}(a,\tau) = \frac{|\tilde{\mathcal{W}}_{x,y}(a,\tau)|^2}{|\tilde{\mathcal{W}}_{x,x}(a,\tau)||\tilde{\mathcal{W}}_{y,y}(a,\tau)|},$$

Based on the extended R^2 coefficient, we can construct a coefficient of determination between two wavelet spectra

$$WER^2_{z,x} = \frac{\int_0^\infty \left(\int_{-\infty}^{\infty} |\tilde{\mathcal{W}}_{z,x}(a,\tau)| d\tau\right)^2 da}{\int_0^\infty \left(\int_{-\infty}^{\infty} |\tilde{\mathcal{W}}_{z,z}(a,\tau)| d\tau \int_{-\infty}^{\infty} |\tilde{\mathcal{W}}_{x,x}(a,\tau)| d\tau\right) da}.$$

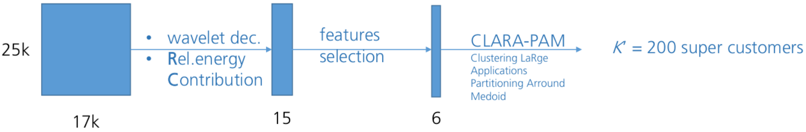
And obtain a dissimilarity [1] based on it

$$d(z,x) = \sqrt{JN(1 - \widehat{WER}^2_{z,x})}$$

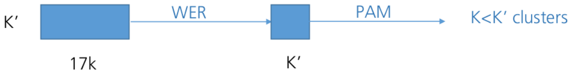
1. A. Antoniadis, X. Brossat, J. Cugliari, and J.-M. Poggi. Clustering functional data using wavelets. International Journal of Wavelets, Multiresolution and Information Processing, 11 :1, 2013.

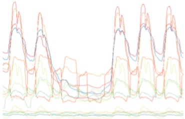
A 2-stages strategy (Energycon)

- 1st stage: create a large number of K' = 200 super customers *fast and scalable*



- 2nd stage: (Ward) ascendant hierarchical clustering of the K' super customers with WER (wavelet coherence) distance *coherent with the forecasting algorithm, computer intensive*





1. J. Cugliari, Y. Goude and J. M. Poggi, "Disaggregated electricity forecasting using wavelet-based clustering of individual consumers," 2016 IEEE International Energy Conference (ENERGYCON), Leuven, 2016, pp. 1-6.

Nonparametric prediction of functional time series

- Let $(Z_k, k \in Z)$ be a stationary sequence of H-valued r.v. Given $Z_1, \ldots, Z_n$ we want to predict the future value of Z_{n+1}.
- A predictor of Z_{n+1} using $Z_1, Z_2, \ldots, Z_n$ is

$$\widetilde{Z_{n+1}} = \mathbb{E}[Z_{n+1} | Z_n, Z_{n-1}, \ldots, Z_1].$$

Autoregressive Hilbertian process of order 1

- The $\textsc{arh}(1)$ centred process states that at each k,

$$Z_k = \rho(Z_{k-1}) + \epsilon_k$$

  where ρ is a compact linear operator and $\{\epsilon_k\}$ an $H-$valued strong white noise. Under mild conditions, this equation has a unique solution which is a strictly stationary process with innovation $\{\epsilon_k\}_{k \in Z}$. [Bosq, (1991)]
- When Z is a zero-mean $\textsc{arh}(1)$ process, the best predictor of Z_{n+1} given $\{Z_1, \ldots, Z_{n-1}\}$ is :

$$\widetilde{Z_{n+1}} = \rho(Z_n).$$

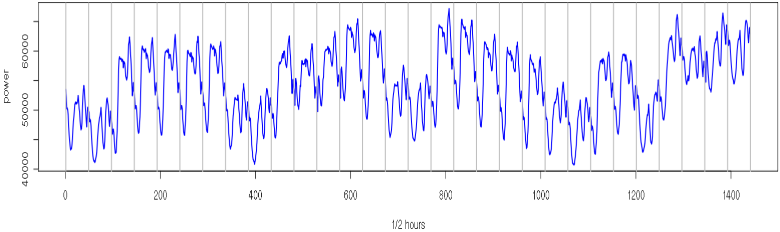
KWF (Kernel Wavelet Functional) in action

KERNEL Estimating directly the prediction equation using the kernel method, the nonparametric estimation is reduced to the direct generalization of the scalar case

KEY the choice of an appropriate distance between current and past situations

IDEA 1 Similar past causes produce similar future consequences.

IDEA 2 Similar shapes form one class.



Prediction of Saturday 10 September 2005

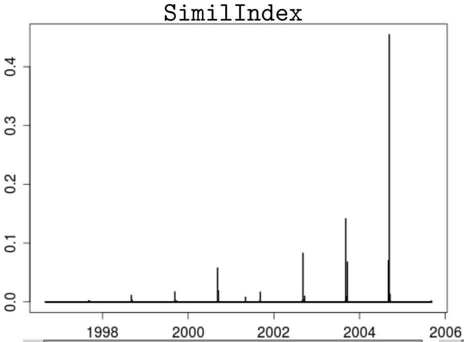
We use Antoniadis *et al.*, (2006) prediction method with corrections to cope with non stationarity Antoniadis *et al.*, (2012, 14).

- The example is the global French electricity load consumption over 10 years
- Use the last observed segment ($n = 9/\text{Sept}/2005$) to look for similar segments in past.
- Construct a similarity index `SimilIndex` (using a kernel).
- Prediction can be written as

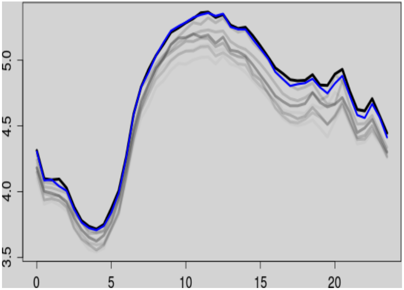
$$\widehat{\text{Load}}_{n+1}(t) = \sum_{m=1}^{n-1} \text{SimilIndex}_{m,n} \times \text{Load}_{m+1}(t)$$

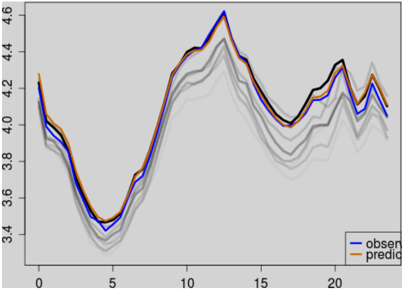
- First difference correction of the approximation part.
- Use of groups to anticipate calendar transitions.
- `R` implementation available in `enercast` package

`github.com/cugliari/enercast`



date	SimilIndex
2004-09-10	0.455
2003-09-05	0.141
2002-09-06	0.083
2004-09-03	0.070
2003-09-19	0.068
2000-09-08	0.058
2000-09-15	0.019
1999-09-10	0.017





past

future

Corrections to handle nonstationarity

On mean level

$$\text{BASE} \quad \widehat{\mathcal{S}_{n+1}(t)} = \sum_{m=1}^{n-1} w_{m,n} \mathcal{S}_{m+1}(t)$$

$$\text{DIFF} \quad \widehat{\mathcal{S}_{n+1}(t)} = \mathcal{S}_n(t) + \sum_{m=2}^{n-1} w_{m,n} \Delta(\mathcal{S}_m)(t)$$

On groups by simple post-processing

Define new weights and renormalize.

$$\tilde{w}_{m,n} = \begin{cases} w_{n,m} & \text{if} \quad gr(m) = gr(n) \\ 0 & \text{otherwise} \end{cases}$$

$gr(n)$ is the group of the n-th segment.

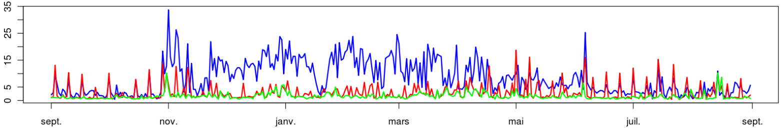


FIGURE : Daily prediction error (in MAPEx100).

As a result of the disaggregation for forecasting

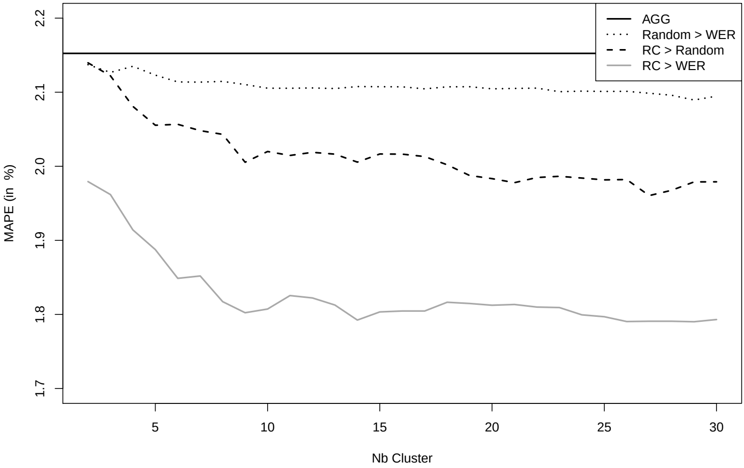


FIGURE : 25K professional EDF clients, sampled at 30min during 5 semesters

How to make it scalable ? the IRSDI project in a nutshell

Context

- Past work in 2014 : clustering with wavelets (RC, WER), KWF, Energycon'16
- Industrial : Electricity demand forecasting & smart grids infrastructure
- Academic : curve's shape & nonparametric function-valued forecast

Aims

- evaluate the upscaling capacity of the Energycon strategy
- if possible, keep the computing centralized

This research benefited from the support of the FMJH 'Program Gaspard Monge for optimization and operations research and their interactions with data science', and from the support from EDF and Thales

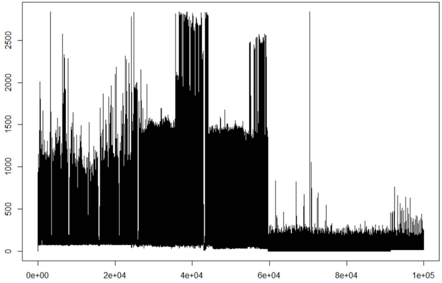
Data description

Available

- EDF : 25K professional clients, sampled at 30min during 5 semesters

Simulated

- using simple simulation scheme we produce datasets of sizes 250K, 2.5M and 25M



Strategies for upscaling

- From 25K to 25M : in 1000 chunks of 25K
- Reference values :
  - $K' = 200$ super consumers (SC)
  - $K^* = 15$ final clusters

1st strategy

- Do 1000 times ONLY Energycon's 1st-step strategy on 25K clients
- With the $1000 \times K'$ SC perform a 2-step run leading to K^* clusters

2nd strategy

- Do 1000 times Energycon's 2-step strategy on 25K clients leading to $1000 \times K^*$ intermediate clusters
- Treat the intermediate clusters as individual curves and perform a single 2-step run to get K^* final clusters

Remark 1 : Mixture models for disaggregated forecasting

- Goal : to theoretically support such a strategy
- Model based clustering using mixture model of functional regressions (Devijver et al. 2016) to build clusters of customers together with the models within each cluster. The model of the kth cluster is

$$Y = \beta_k X + \varepsilon_k$$

  where ε_k is a centered Gaussian noise of covariance matrix Σ_k.
- Let π_k be the proportions of each cluster, and $\xi = (\boldsymbol{\pi}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$ the vector containing all the parameters to estimate. Then, the conditional density is (φ the Gaussian density)

$$s_\xi^K(y|x) = \sum_{k=1}^{K} \pi_k \varphi(\beta_k x, \Sigma_k)$$

- Tools : Multivariate and functional extensions of existing results about sparse estimation of such mixture models + slope heuristic penalized log-likelihood $-\frac{1}{n}\sum_{i=1}^{n} \log(s_\xi^K(y_i|x_i)) + \kappa D_K$

1. E. Devijver, Y. Goude and J. M. Poggi (2016) "Clustering Electricity Consumers using High Dimensional Regression Mixture Models" Preprint, 23 pages, arxiv.org/abs/1507.00167

Remark 2 : Random Forests and Sequential Aggregation for Disaggregated Demand Forecasting

- Goal : to disaggregate in space the global signal to improve forecasts
- Goal 2 : to use a flexible nonparametric forecasting method allowing easily to take into account exogenous variables
- Goal 3 : to generate many models from clusters and combine them
- Method : Random Forests as a forecasting method together with unsupervised clusters and intensive use of sequential aggregation of experts related to the clusters (Goehry et al. 2017, 2018)
- Some partial conclusion on Irish data
  - based on exogenous individual variables, disaggregation leads to a real gain but not more than random
  - Random forests provide useful predictors for all aggregation scales
  - the decisive additional gains are obtained thanks to the combination of cluster's predictors

---

1. B. Goehry, Y. Goude, P. Massart and J. M. Poggi (2018) , "Forêts aléatoires pour la prévision à plusieurs échelles de consommations électriques" *Proc. Journées de Statistique JDS 2018, 28 mai-1 juin, Paris-Saclay, France, talk 112* `toltex.u-ga.fr/users/RCqls/Workshop/jds2018/resumesLongs/subm112.pdf`

More information

- Article freely available (including references and self-references herein)

  Auder, Cugliari, Goude, Poggi (2018)

  Scalable Clustering of Individual Electrical Curves for Profiling and Bottom-Up Forecasting

  in energies vol. 11(7), 1893

  `https://doi.org/10.3390/en11071893`

- R package available from GitHub

  `github.com/cugliari/iecclust`

- Recall that KWF is available from the `enercast` R package at

  `github.com/cugliari/enercast`