Statistical and machine learning methods to model and forecast Energy

Mathilde Mougeot

ENSIIE & LPMA

NUS-USPC
November 2017

Electricity Framework

Production	Consumption
Nuclear power plants	industrial plants
coal fired power plants	home and heating
wind farms	building heating
photovoltaic farms	...

$\rightarrow$ Electricity can hardly be stored. There is a need to :
Balance between electrical production and consumption
Forecast consumption and production

Statistical and machine learning methods to model and forecast Energy

We have studied models for Energy in several directions :

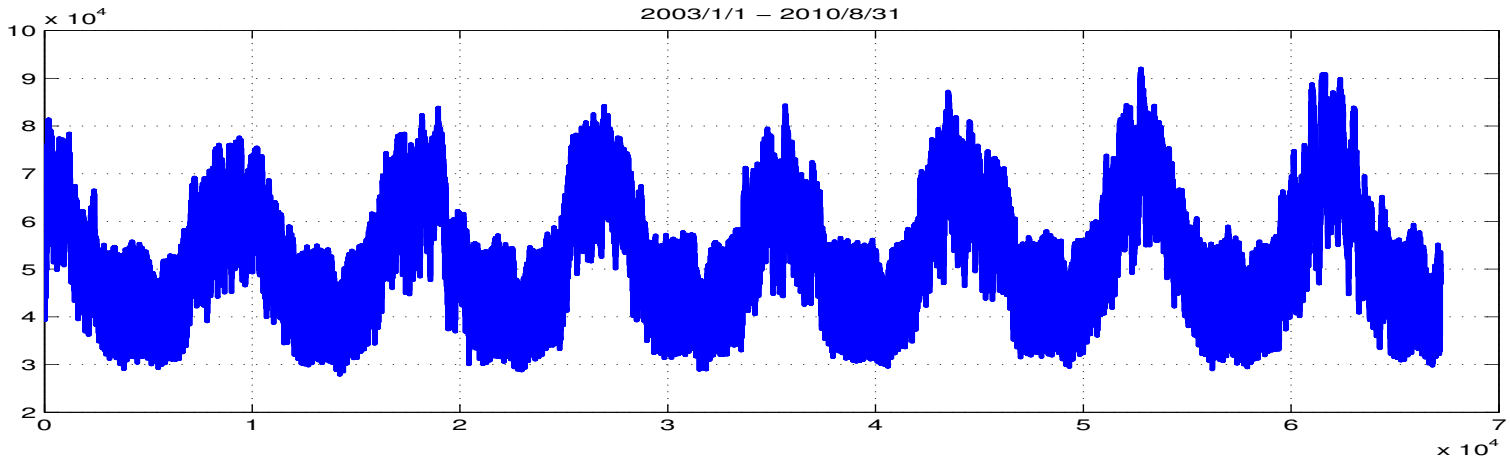
- Consumption
  High dimensional regression models to Forecast the French National Consumption. with D. Picard, K. Tribouley, V. Lefieux (RTE), JRSSB, AADA

- Production
  Modeling the Wind Farm production using machine learning tools with A. Fischer, L. Montuelle, D. Picard, Wind Energy 2017

- Savings
  Over consumption monitoring and diagnostic for industrial equipments : aggregation with O. Cadet (Air Liquide)

Electricity Production

Motivation : balance between electrical production and consumption



Electrical Consumption Time series

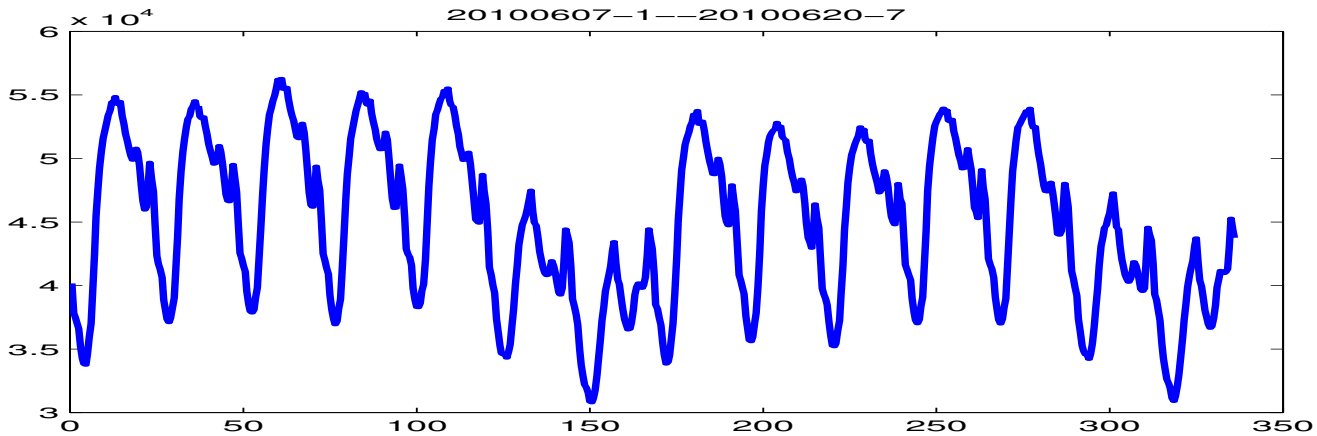


Figure – Two weeks of The French National electrical consumption

RTE requirement :

"Is it possible to built forecast models in the electricity consumption field which would rely on very few parameters and would be easy to calibrate without the need of human expertise - and which at the same time, would show good performances ?"

RTE Current operational model

Today, RTE provides day-ahead load forecasting using

- A statistical forecast model METEHORE (decision-making tool)
- an a human expertise ("The forecasters")

Remark : The metehore model provides accurate forecast. However, it depends on a very large number of parameters and is no enough adaptive, whereas the electricity demand is evolving in a changing context

The context today :

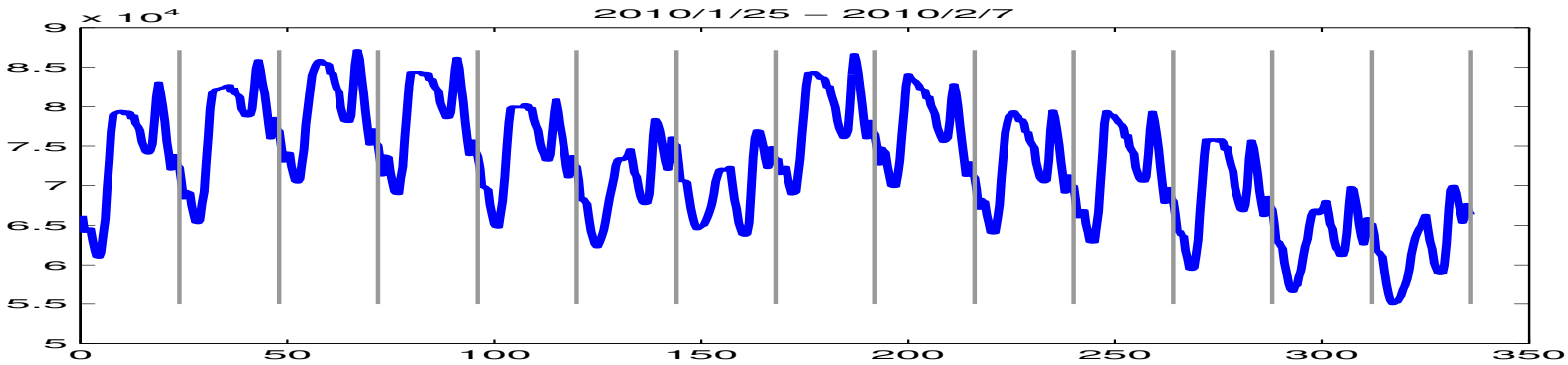
- The evolution of electrical uses, the energy-demand management, the smartgrids context... induce changes in consumers' behavior
- New electric heating systems (such as heat pumps) induces variations in the weather-sensitivity of consumption
- The development of decentralized production (wind, solar) has an indirect impact on the consumption signal as the electrical transport network nodes

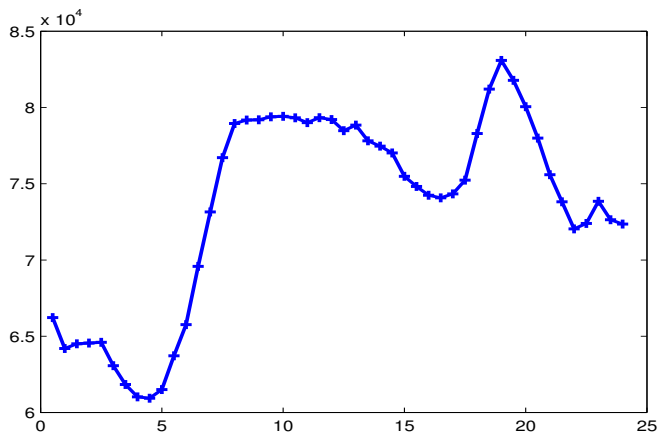
  $\rightarrow$ Forecast models of consumption have to adapt to quick changes in the load curve and to better control large errors

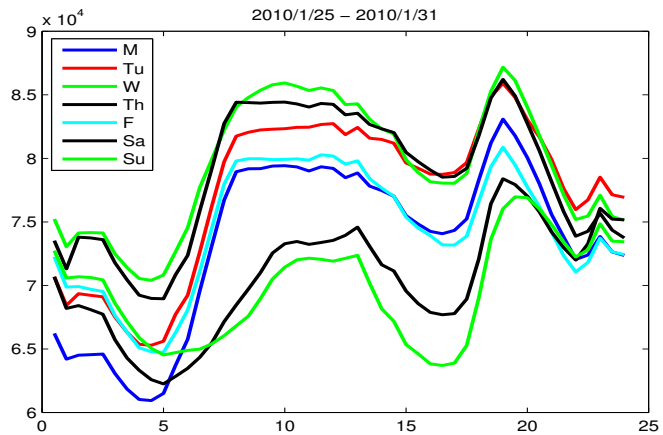
Recent works and existing models

- Time series analysis : exponential weights (Taylor 2012)
- Non parametric regression (J.M. Poggi 1994, Antoniadis et al. 2006, J. Cugliari 2011)
- Regression tree (Y. Goude et al. 2013,..)
- Aggregation (P. Gaillard & Y. Goude 2014)
- Model-based clustering (E. Devijver 2014)

Intraday load curve :Fonctional data



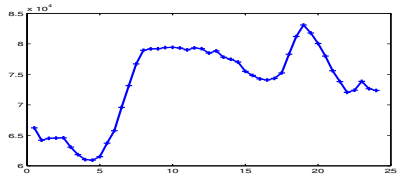




Intra day load curve, 30' sampling (48 pts),
$Y \in \mathbb{R}^{n=48}$ (Y_t $1 \leq t \leq 2800$)

Modeling each intra day signal as a function

We investigate the problem in a supervised learning setting :



- We consider each time unit signal :

$$Z_i = (Y_i, U_i),\ i = 1, \ldots, n = 48$$

- For each signal, we want to identify f, an unknown function such that :

$$Y_i = f(U_i) + \epsilon_i.$$

where :

  - The generic consumption signal observed on the time unit :

$$Y_i,\ i = 1, \ldots, n$$

  - The design (here fixed equi distributed) : $U_i = \frac{i}{n}$

Using a dictionary

Consider a dictionary $\mathcal{D}$ of functions $\mathcal{D} = \{g_1, \dots g_p\}$ and
Assume that f can be well fitted by this dictionary

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

where h is a 'small' function (in absolute value).
The model is

$$Y_i = \sum_{\ell=1}^{p} \beta_\ell g_\ell(U_i) + h(U_i) + \epsilon_i', \ i = 1, \dots, n$$

which coincides with the linear model :
$Y = X\beta + \epsilon$ *with* $X(n \times p)$ putting $\epsilon_i = h(U_i) + \epsilon_i'$ and
$G_{i\ell} = g_\ell(U_i)$.

High dimensional framework

Solution : $\hat{\beta} = Argmin||Y - X\beta||^2$

- More variables (functions) than observations $n << p$

$$\begin{bmatrix} y_1 \\ y_2 \\ \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & \dots & x_{1p} \\ & & & \\ & & & \\ x_{n1} & & \dots & x_{np} \end{bmatrix} * \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \dots \\ \\ \beta_p \end{bmatrix} + \epsilon$$

"Fat matrix"

$\rightarrow$ Infinity of $\hat{\beta}$ solutions.

$\rightarrow$ Need more assumptions on β to solve the problem

$\rightarrow$ Ex : Lasso (ℓ_1 penalization), Ridge (ℓ_2)...

Theoretical background : Learning Out of Leaders

- $Y = X\beta + \epsilon$, $\epsilon \sim \mathcal{N}(0, \sigma^2)$, β unknown
- $\hat{\beta} = Argmin||Y - X\beta||^2$, OLS

Sparse approximation using Thresholding : Learning Out of Leaders* :

- based on 2 Thresholding steps,
- weak complexity, sparse and non linear solution,
- Algorithm in 3 steps (X column normalized, $\sum_j X_j^2/n = 1$) :
- Concistency results

step		compute	size
1. SELECTION (threshold λ_1)	Find b Leaders $b < n << p$	X_b	(n, b)
2. REGRESSION	on Leaders	$\tilde{\beta} = (X_b^T X_b)^{-1} X_b^T Y$	$(1, b)$
3. THRESHOLD λ_2	the coefficients	$\hat{\beta}$	$(1, \hat{S})$

(*) *MM, D.*

Picard, K. Tribouley, JRSS B 2012,B Stat. Methodol. vol 74

LOL assumptions and thresholds

- When :
  1. **Sparsity :**
     $B_0(S,M) := \{\beta \in \mathbb{R}^p,\ \sum_{j=1}^p I\{|\beta_j \neq 0\} \leq S,\ \|\beta\|_{l^1(p)} \leq M\}$.
  2. **Dimension :** $p \leq \exp(\square n)$,
  3. **Coherence :** $\tau_n \leq \square\sqrt{\frac{\log p}{n}}$ ("max of correlation between columns")
- Choose : **the thresholds** λ_1**,** λ_2

  $\lambda_1 = \square\sqrt{\frac{\log p}{n}},\ \lambda_2 = \square\sqrt{\frac{\log p}{n}}$
- Approximation, Concentration results :
  - Prediction loss : $\frac{1}{n}\sum_{i=1}^n(\widehat{Y}_i - \mathbb{E}Y_i)^2 = d(\hat{\beta}^*,\beta)^2$

$$\sup_{\beta\in B_0(S,M)} \mathbb{P}\left(d(\hat{\beta}^*,\beta) > \eta\right) \leq \begin{cases} 4e^{-\gamma n\eta^2} & \text{for} \quad \eta^2 \geq DS[\sqrt{\frac{\log p}{n}} \vee \tau_n]^2 \\ 1 & \text{for} \quad \eta^2 \leq DS[\sqrt{\frac{\log p}{n}} \vee \tau_n]^2 \end{cases}$$

(*) *MM, D. Picard, K. Tribouley, JRSS B 2012,B Stat. Methodol. vol 74*

Generic Dictionary

- Each day t, $\boxed{Y_t = X\beta_t + \epsilon_t}$
- with Dictionary of p functions $\mathcal{D} = \{g_1, \dots g_p\}$ $G_{i\ell} = g_\ell(U_i)$
- For daily load curves ($dim(Y_t = 48)$:
  a good choice happened finally to be a mixture of the Fourier basis and the Haar basis, $p = 62$.
  1. (1 :1) constant function (1)
  2. (2 :24) cosine functions (with increasing frequencies) (23)
  3. (25 :47) sine functions (with increasing frequencies)(23)
  4. (48 :62) Haar functions (with increasing frequencies)(15)
- Approximation : $p = 7$, $E_{MAPE} = 1.4\%$

June 17^{th}, 2009

$S = 5$, $MAPE = 0.0147$

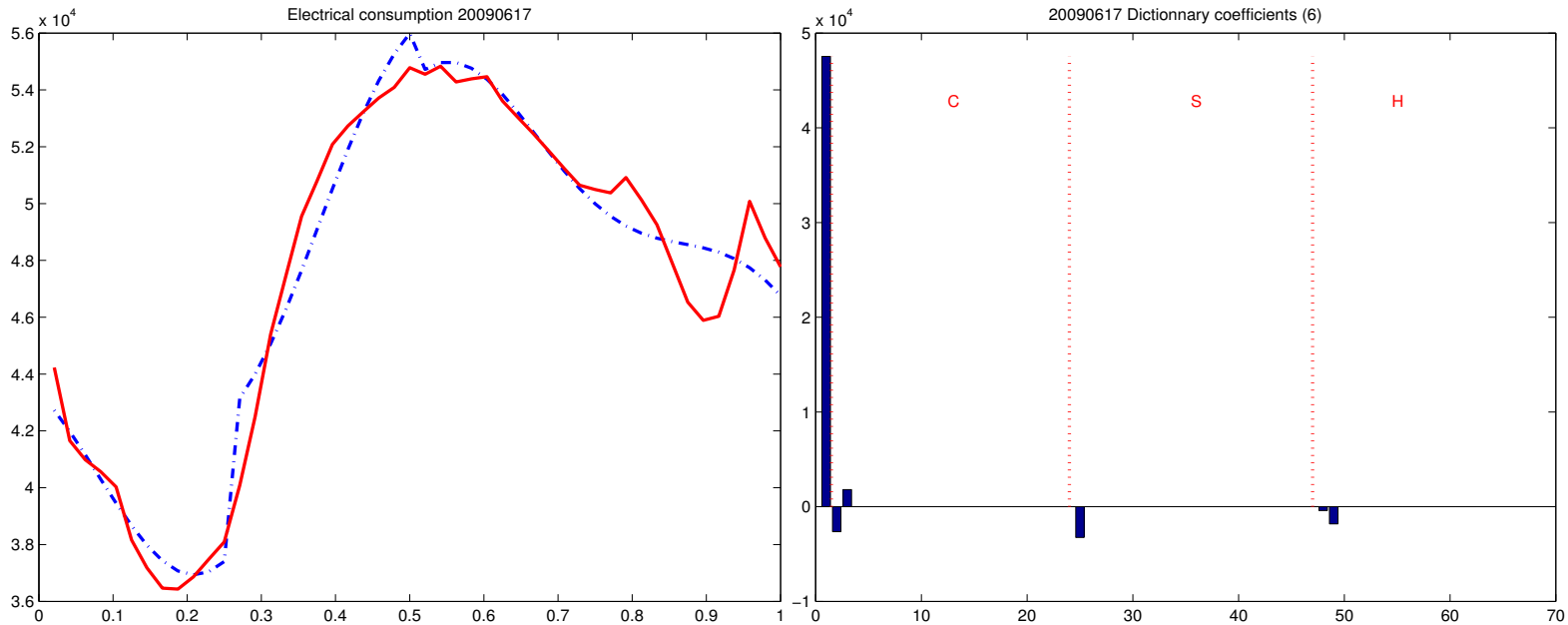


Figure – 2003 04 30

left : observed signal - red line, approximated signal -blue line

right : S coefficients on the dictionary

November 18^{th} 2007

$S = 12$, $MAPE = 0.0057 = 0,57\%$. $MAPE = \frac{1}{n}\sum_{i}^{n=48} |Y_i - \hat{Y}_i|/Y_i$

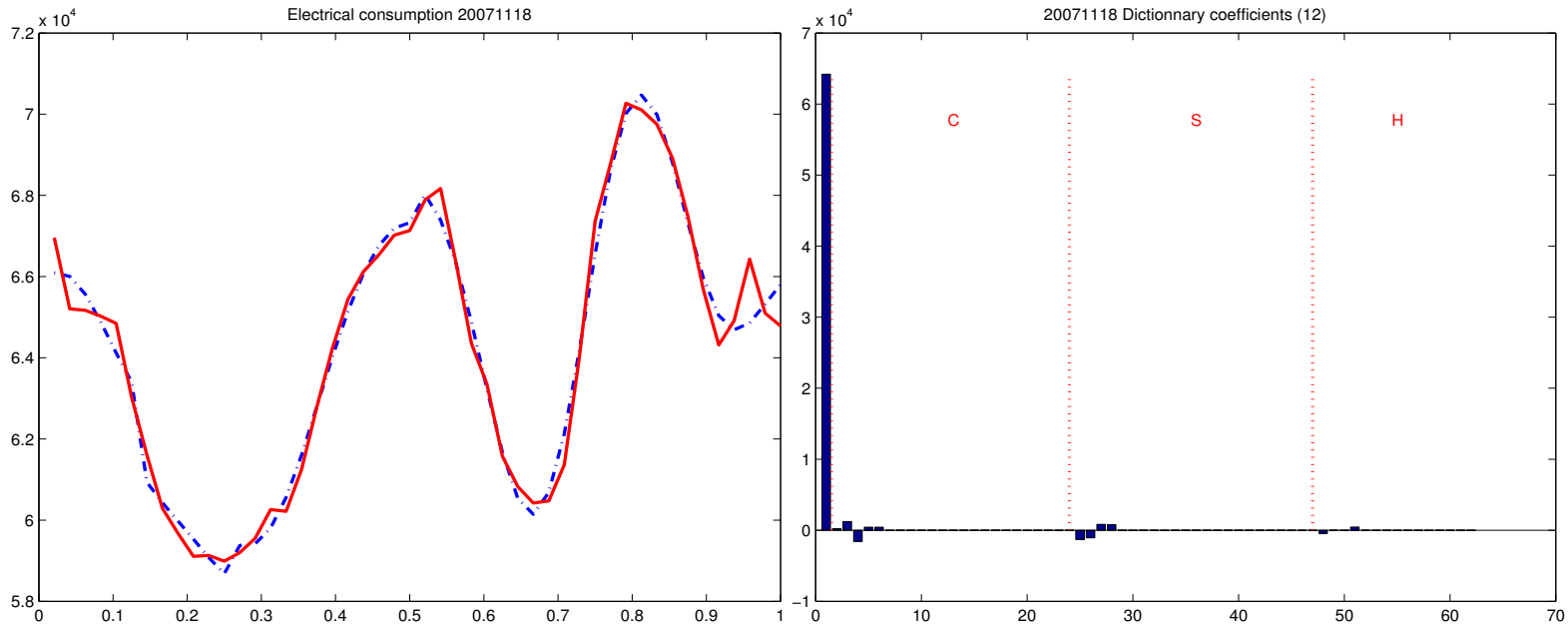


Figure – 2007 11 18

left : observed signal - red line, approximated signal -blue line

right : S coefficients on the dictionary

Spot of Temperatures, Cloud Cover and Wind information

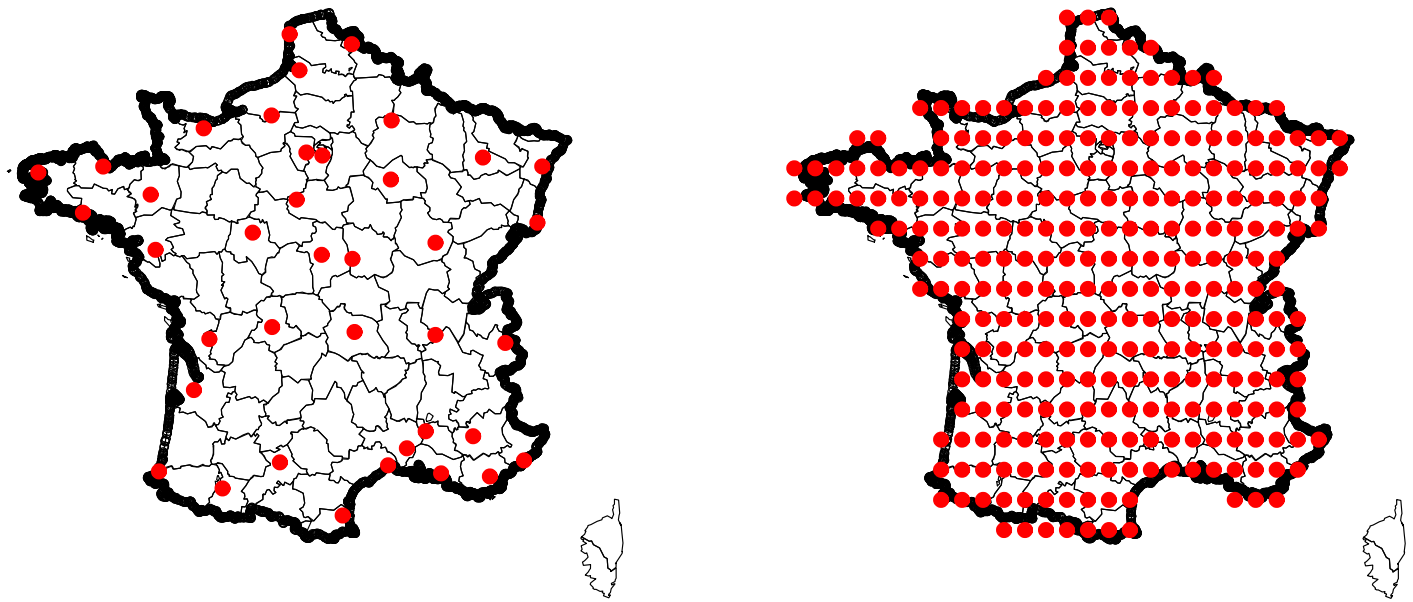


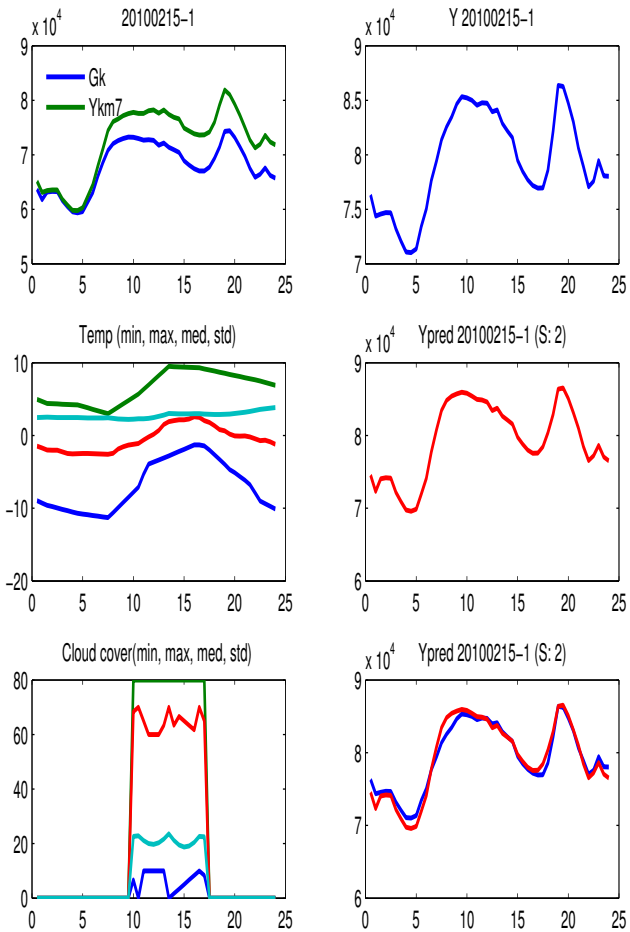
Figure – Temp., Cloud Cover spots (♯39) and wind data (♯293)

Intraday Specific Dictionary

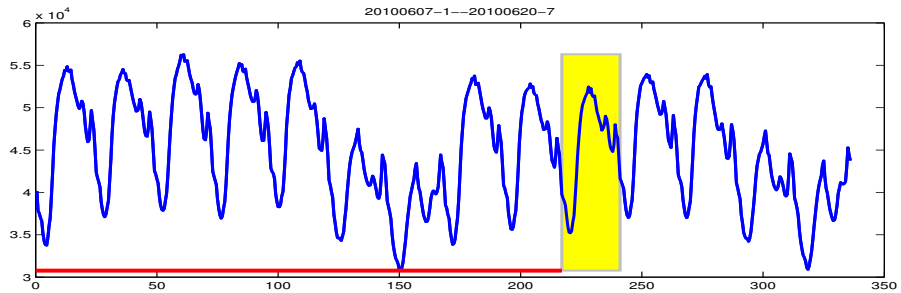
- Each day t, $\boxed{Y_t = X_t \beta_t + \epsilon_t}$
- with Dictionary of p functions $\mathcal{D}_t = \{g_1^t, \dots g_p^t\}$
  Final model, $p = 10$ ($p = 14$)
  1. **2**, Shape fonctions (group centroid, previous week day Y_{t-7})
  2. **8**, Climate fonctions (Temperature and Cloud Cover Indicators computed over the 39 meteorological spots. (and Wind...(+4))
- Approximation performance :
  - LOL adaptive using shape and meteorological variables
    - $S = 2.35$ [2 ;6],
    - $\bar{E}_{MAPE} = 1.5\%$
  - LOL adaptive using a generic dictionary
    - Trigonometric-Fourier
    - $S = 7$
    - $\bar{E}_{MAPE} = 1.7\%$

Illustration : intraday load curve model/prediction (Winter Monday, Thursday)

$$Y_t = X_t \beta_t + \epsilon_t$$



From Sparse approximation to Forecast



Each day t :

- $Y_t = \hat{Y}_t + \hat{\epsilon}_t$ with model : $\hat{Y}_t = \sum_{j=1}^{p} \hat{\beta}_j^t g_j^t$
- Forecast $(d > t)$ $\tilde{Y}_d = \sum_{j=1}^{p} \tilde{\beta}_j^d g_j^d + \delta_{\mathbf{d}}$

Looking for a good candidate of coefficients in the past :

- Plug in estimated coefficients
- $\tilde{\beta}_d = \hat{\beta}_{\mathcal{M}(d)}$ with $\mathcal{M}(d) << d$
- $\mathcal{M}$ "Expert"

Expert $\mathcal{M}$ to forecast

Strategy Let $\mathcal{M}$ be a function (strategy), from $\mathbb{N}$ to $\mathbb{N}$ such that for any $t \in \mathbb{N}, \mathcal{M}(t) < t$. (data dependent or not)

Plug-in To the strategy $\mathcal{M}$ we associate the expert $\tilde{Y}_d^{\mathcal{M}}$: the forecast of the signal of day d using prediction strategy $\mathcal{M}$.

$$\tilde{Y}_d^{\mathcal{M}} = \sum_{j=1}^{p} \hat{\beta}^j_{\mathcal{M}(d)} g_d^j + \delta_d$$

$\hat{\beta}^j_{\mathcal{M}(d)}$, $j = 1, \ldots, p$ are the estimated coefficients computed with LOL algorithm at day $\mathcal{M}(d)$.

Specialized Experts focus on

Nearest neighbor strategies based on different variables and metrics :

1. (2) Time depending (t-1, t-7)
2. (2) climatic configuration of the day (Temperature)
3. (2) constrained climatic configuration of the day (Temperature/Cloud Covering)
4. group constraint climatic configuration of the day (Temperature/group)
5. climatic configuration of the day constrained by the type of the day (Temperature/day)
6. climatic configuration of the day constrained by a calendar group (Temperature/calendar)
7. climatic configuration of the day (Cloud cover)
8. group constraint climatic configuration of the day (Cloud Covering/group)
9. climatic configuration of the day constrained by the type of the day (Cloud Covering/day)
10. climatic configuration of the day constrained by a calendar group (Cloud Covering/calendar)
11. Wind ...

MAPE Forecast performances

Forecast results are computed using one year of data from 1^{st} September 2009 to 31^{th} August 2010.

M	mean	med	min	max
Naive	0.0634	0.0415	0.0046	0.1982
Apx	0.0183	0.0151	0.0035	0.0862
tm1	0.0323	0.0262	0.0050	0.1412
tm7	0.0303	0.0239	0.0056	0.1920
T	0.0305	0.0242	0.0065	0.2232
Tm	0.0321	0.0264	0.0062	0.2138
T/N	0.0328	0.0258	0.0043	0.4762
Tm/N	0.0321	0.0248	0.0057	0.1639
T/G	0.0337	0.0247	0.0058	0.4762
T/d	0.0330	0.0257	0.0052	0.3749
T/c	0.0314	0.0249	0.0054	0.1848
Cs/G	0.0297	0.0230	0.0047	0.1915
C/d	0.0281	0.0219	0.0036	0.2722
C/c	0.0288	0.0224	0.0036	0.2722

Aggregation of predictors : Exponential weights

$$\tilde{Y}_d^{wgt*} = \frac{\sum_{m=1}^{M} w_d^m \tilde{Y}_d^m}{\sum_{m=1}^{M} w_d^m}$$

with

$$w_d^{\mathcal{M}} = exp(-|\hat{Y}_{d^*_{\mathcal{M}}} - Y_{d^*_{\mathcal{M}}}|_2^2/\theta)$$

θ is a parameter, calibrated by cross-validation.

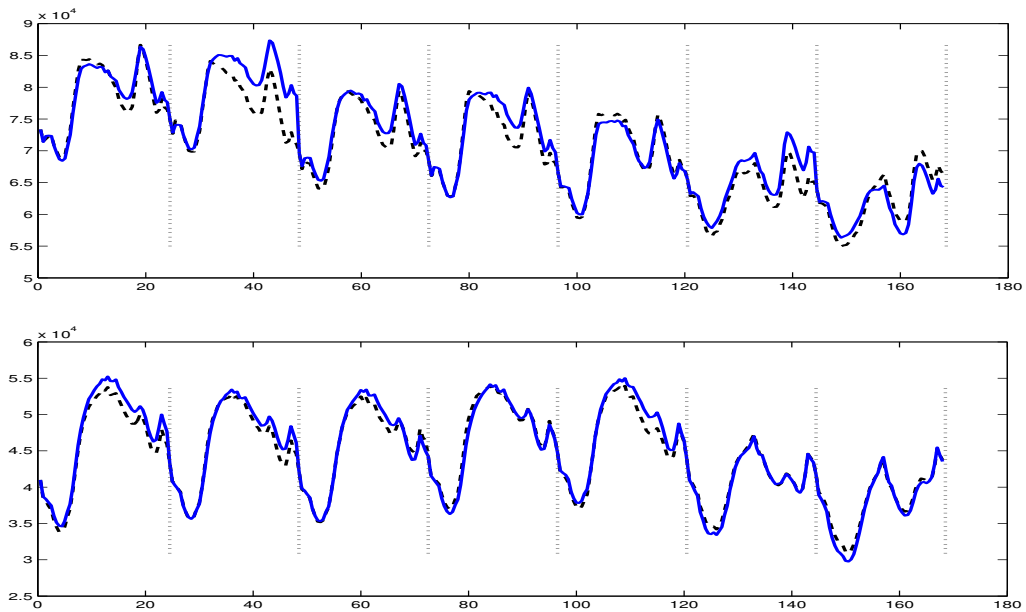
Performances after aggregation

Mape performances for aggregated methods computed for one year

mean	med	min	max
0.0230	0.0197	0.0052	0.0695

Mape performances for Oracle computed for one year

mean	med	min	max
0.0144	-	-	0.074



Conclusion

- Competitive approach compared to usual time serie analysis with much less parameters.
- Sparse approximation
  - a Generic dictionary for compression and pattern extraction
  - Intra day specific dictionaries for approximation and prediction
- Forecasting
  - Various experts for prediction based on a retrieval information strategy
  - Aggregation using exponential weights,
- FOREWER project research
  - prediction for renewable energy with machine learning methods
- work in progress for improvement