Data Stream and Load Forecasting :

Some ideas for a research project at Electricité De France :

Alain DESSERTAINE

Outline

1. Context
   - *Some words about the potential use of the data Streams at EDF*
   - *The electric consumption forecasting: Why and how?*
2. The first reflections and some tracks of research
   - Some remarks about the source data
   - Some remarks on streams and summaries of data
   - Some preliminary ideas for the exploitation of the data streams for forecasting models

1

Context

1. Context

➢ *Some words about the potential use of the data Streams at EDF*

➢ **The use of the data coming from the 32 millions of communicating meters (installation planed from now to 2013)**

- ➢ Queries :
  - ➢ Allocation of global consumption among all operator in « real time »
  - ➢ Information system of the Electricity consumption
- ➢ Analysis and modeling :
  - ➢ Marketing and commercial analysis of uses
  - ➢ Load Forecasting

1. Context

The electric consumption forecasting: Why?

- ***Long-term (from 2 to more than 20 years)***
  ⇔ *Investments in order to make the park of production*

- ***Middle-term (from 13 days to 2 years)***
  ⇔ *Adaptation of the production equipments and the potential purchase on the financial markets*

- ***Short-term (from one hour to 12 days)***
  ⇔ *Balance between Consumption / Production and Sourcing*

1. Context

The electric consumption forecasting: how?

- *Aggregate time series modeling*
  - *Impossibility to compute, in real time, « official » time series, especially for costumers which have not telemeters meters (method based by profiling methods)*

- *Non-linear additive models :*
  - *A part dependent on the climate (mostly temperature, and for some models, cloud cover)*
  - *A part embedding seasonality and trend.*

- Aggregate forecasting test:
  - Building different models on « natural » wallets, and summarizing of the forecasts
  - Same thing, but on groups chosen (or built) specifically to improve global forecast (by clustering or classication methods)

2

Short-term forecasting and data-Stream : The first reflections and some tracks of research

2. The first reflections and some tracks of research

- Some remarks about the source data :
  - Space and time characteristics
  - "quasi-continuous" temporal characteristics
    - Minute by minute measurement (perhaps second by second?)
  - Information and data by uses
  - Internal and/or external temperature measures

Exhaustive recuperation, or Space and time sampling?

2. The first reflections and some tracks of research

➢Some remarks on streams and summaries of data :

- We will need to establish quite long historical curves in order to collect the temporal phenomena such as the tendencies, the seasonal variations and other periodic phenomena of our data
  - the temporal granularities could be all the more broad as the stored data will be distant in the past
- We will need to work with sufficiently aggregate levels because aggregate curves would be less random or erratic
- We will need to connect the collected curves to reliable data allowing the qualification of our data :
  - geographical data (even socio-economic)
  - contractual data (even information on the uses)
  - weather data: measurements, even forecast data

2. The first reflections and some tracks of research

➢Some hypothesis constraints :

- We will not be able to preserve (nor to collect) the whole of the curves of consumption at the step second or the step minute for all customers .
  - ***Management of a panel with a great number of customers, and that we will work on specific summaries of theirs consumption curves, according to the complexities of their own process of consumption .***
- collect curves with different, but constant temporal granularities?
- if the granularity can be established on the level of the meter, maybe we will be able to recover curves with variable granularities in time for each sampling curve in order to collect some increasing, decreasing or important varying individual consumption periods.
  - But, how could we summarizing those information? (with a curves census, or with a curves sampling)

We just have now to think about the way to use these restitutions of data streams.

2. The first reflections and some tracks of research

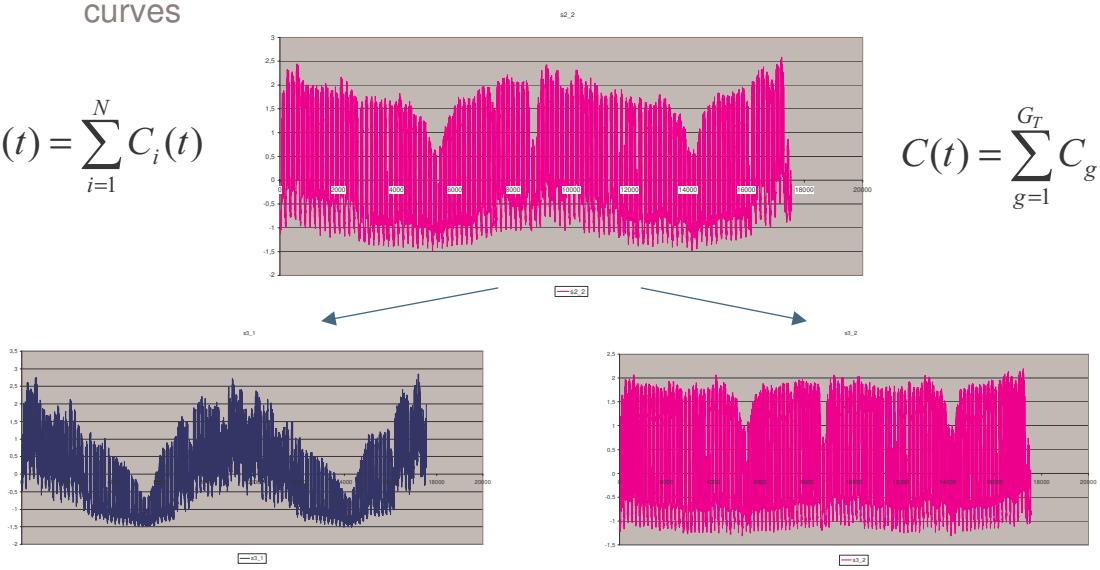
➢Some preliminary ideas for the exploitation of the data streams for forecasting models

- ➢ Clustering or classification to forecast by aggregation/disintegration of curves
- ➢ Some ideas for the forecasting models and incremental approaches
- ➢ And forecasting models in an environment of data stream ?

2. Some preliminary ideas for the exploitation of the data streams for forecasting models

- ➢ Clustering or classification to forecast by aggregation/disintegration of curves

$$C(t) = \sum_{i=1}^{N} C_i(t)$$



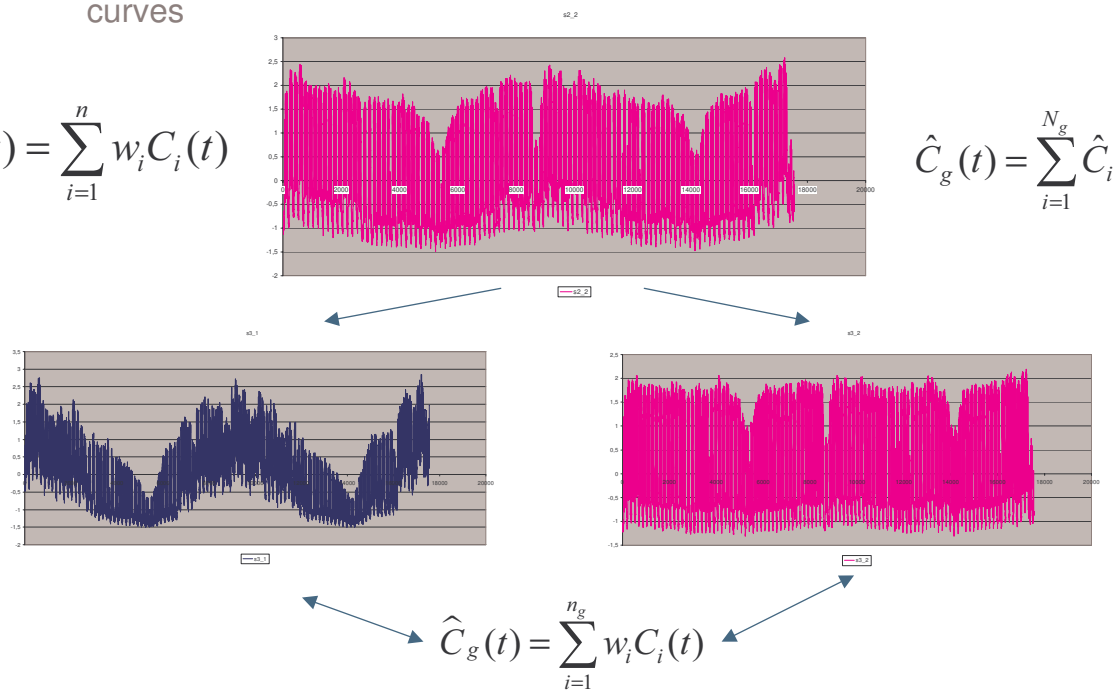
$$C(t) = \sum_{g=1}^{G_T} C_g(t)$$

$$C_g(t) = \sum_{i=1}^{N_g} C_i(t)$$

2. Some preliminary ideas for the exploitation of the data streams for forecasting models

- Clustering or classification to forecast by aggregation/disintegration of curves

$$\hat{C}(t) = \sum_{i=1}^{n} w_i C_i(t)$$



$$\hat{C}_g(t) = \sum_{i=1}^{N_g} \hat{C}_i(t)$$

$$\hat{C}_g(t) = \sum_{i=1}^{n_g} w_i C_i(t)$$

2. Some preliminary ideas for the exploitation of the data streams for forecasting models

- **Aggregation / Disaggregation of curves and « Abundance principle » :**
  - for an individual i , element of a group g (with g ∈ G_T, partition of the population at a period T), its signal of consumption could be divided in an additive way into two random signals :

$$C_i(t) = \lambda_i S_g(t) + \xi_i(t)$$

  - Now let' us going on with the hypothesis of independence over the time of each two distinctive signals of a group g :

$$Cov_t(\xi_i(t), \xi_j(t)) = 0 \quad \forall (i,j) \in g$$

  - With :

$$Cov_t(\xi_i(t), \xi_j(t)) = \int_T \left(\xi_i(t) - \frac{1}{T} \int_0^T \xi_i(k).dk \right) \left(\xi_j(t) - \frac{1}{T} \int_0^T \xi_j(k).dk \right)$$

2. Some preliminary ideas for the exploitation of the data streams for forecasting models

➢ Aggregation / Disaggregation of curves and « Abundance principle » :

- The consumption global signal of a group g :

$$C_g(t) = S_g(t)\sum_{i=1}^{N_g}\lambda_i + o(S_g(t)\sum_{i=1}^{N_g}\lambda_i)$$

- In a sampling case, we have :

$$\hat{C}_g(t) = \hat{S}_g(t)\sum_{i=1}^{n_g} w_i\lambda_i + o(\hat{S}_g(t)\sum_{i=1}^{N_g} w_i\lambda_i)$$

2. Some preliminary ideas for the exploitation of the data streams for forecasting models

➢ Aggregation / Disaggregation of curves and « Abundance principles » :

- Let us suppose that we can build models “perfectly adapted” to each signal $S_g(t)$, a disintegration of our global signal could result in the fact that the composite predictor of the predictors built on these G models independent are at any moment (or on average over a given period) more powerful than “the best” predictor built on the knowledge of the global signal C(t) and of the exogenous variables. :

$$\bar{Q}(\sum_{g=1}^{G_T}\hat{S}_g(t)\sum_{i=1}^{N_g}\lambda_i, t\in p) \le \bar{Q}(\hat{C}(t), t\in p)$$

- In a sampling case, we have : $\hat{S}_g(t)\sum_{i=1}^{n_g} w_i\lambda_i$

2. Some preliminary ideas for the exploitation of the data streams for forecasting models

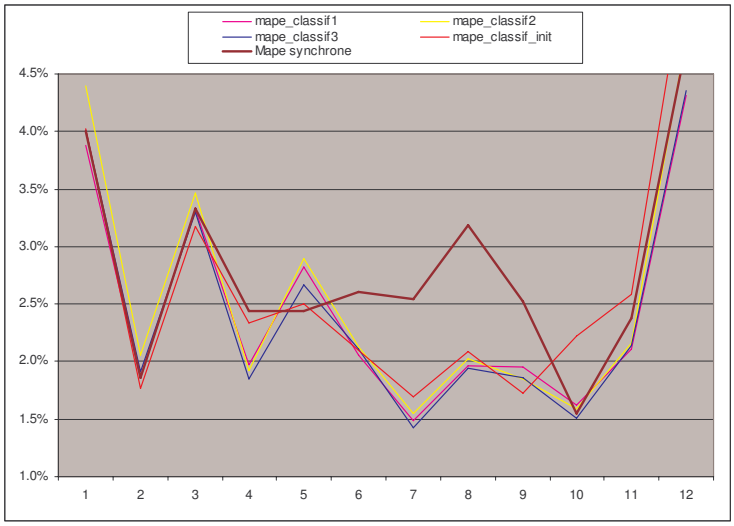
➢ **Aggregation / Disaggregation of curves and « Abundance principles » :**

➢ Then, we can say that the forecasting quality will depend of :

- ➢ Appropriateness of the partitions to build
- ➢ The quality of the models used
- ➢ The sampling and adjustment errors

2. Some preliminary ideas for the exploitation of the data streams for forecasting models

Appropriateness of the partitions to build



-> We must to work yet with a new distance to include the predictability notion in the clustering or classification process

-> How can we update partition GT? Stream-mining?

2. Some preliminary ideas for the exploitation of the data streams for forecasting models

➢ **Aggregation / Disaggregation of curves and « Abundance principles » :**

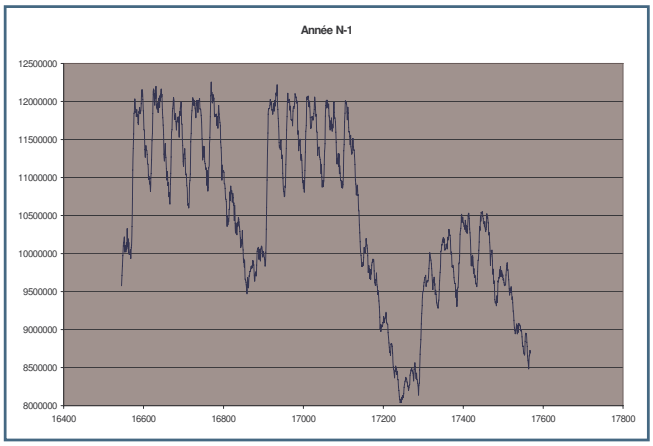
- The data will be very fine temporal and variable by the time granularities
  - *Hilbertian Auto-regressif models (Bosq)*
  - *Non parametric models on functional decomposition (for example : additive models with similarity on wavelet coefficients (Cf. Antoniadis)*
- The data will be of space and time nature, and will have to undoubtedly remain it at the end of the phases of classification (because of the space aspect in classification).
- To work currently will enable us to use recent data on the level of each studied wallet.
- Use Symbolic data models to take account the sampling error in each measure of the estimate curve)

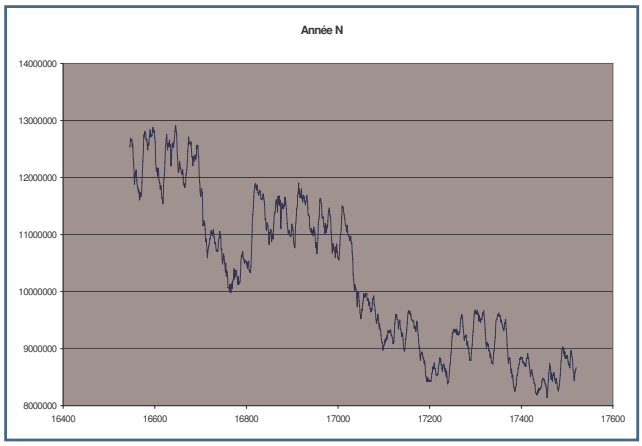
2. Some preliminary ideas for the exploitation of the data streams for forecasting models

➢ **Aggregation / Disaggregation of curves and « Abundance principles » :**

- Balancing sampling and / or calibration estimation on functional decomposition (Wavelets, Non-linear decomposition with wavelets, or using firsts functional principal components …)
- First trials with calibration estimators using simple wavelet decomposition :

2. Some preliminary ideas for the exploitation of the data streams for forecasting models





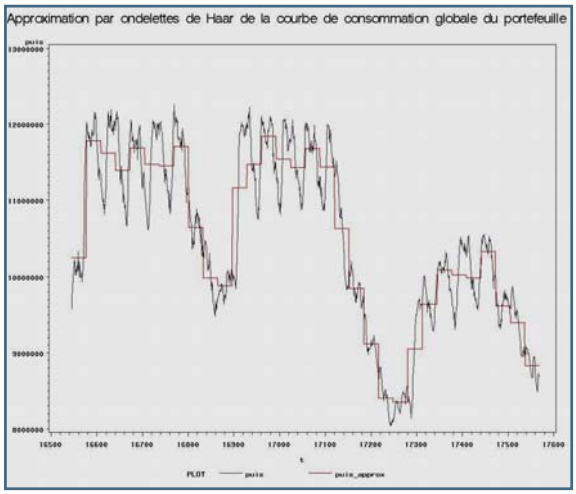
➢With an unequal probabilities curve sampling, we estimate the total curve with with sampling design and curves of the sample for the year N!

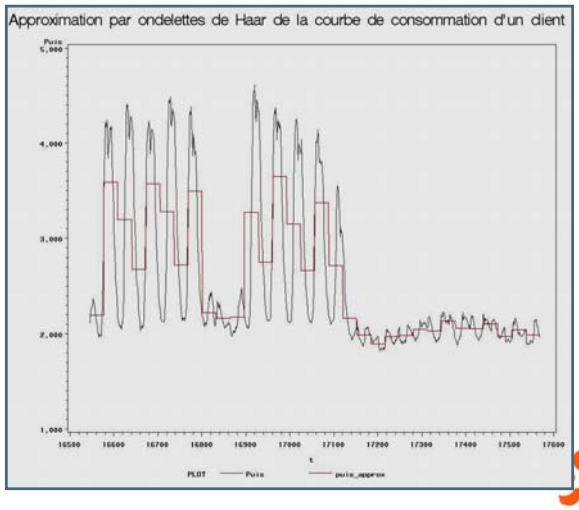
➢To try to improve the result, we use the same data, but in the year N-1, and the global serie known of the wallet. We compute new weighting coefficients to estimate the real curve we know!

2. Some preliminary ideas for the exploitation of the data streams for forecasting models

- We compute the wavelet decomposition for the sample and global curves in the year N-1 :

$$X^{(i)} = A_J^{(i)} + \sum_{j=1}^{J} D_j^{(i)}$$





2. Some preliminary ideas for the exploitation of the data streams for forecasting models

- We know that :

$$A_J^{TOT} = \sum_i A_J^{(i)} \qquad et \quad D_J^{TOT} = \sum_i D_j^{(i)} \setminus \quad \forall j \in [1, J]$$

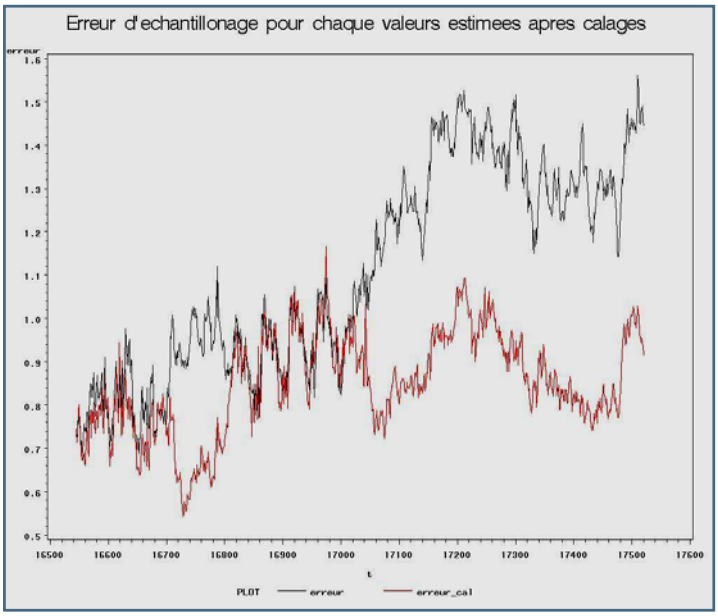
- Then, we can estimate each coefficient $A_{k,5}^{TOT}$ with similar coefficients of the sample curves $A_{k,5}^{(i)}$ and with their unequal probability π_i :

$$\hat{A}_{k,5}^{TOT} = \sum_{i \in S} \frac{A_{k,5}^{(i)}}{\pi_i}, \quad \forall k \in \{1, ..., 32\}$$

- We use the macro CALMAR to compute new weights :

$$A_{k,5}^{TOT} = \sum_{i \in S} \omega_i A_{k,5}^{(i)}, \quad \forall k \in \{1, ..., 32\}$$

2. Some preliminary ideas for the exploitation of the data streams for forecasting models



-> temporal dependences weighting ? (Calibration using a Kalman – Filter under contraints?…)

2. Some preliminary ideas for the exploitation of the data streams for forecasting models

Then, what about load forecasting using data Stream?

- First, we can imagine some adapted construction of stream summarize or stream-mining objects or models?
  - Classification? And incremental's classification using data-stream
  - Global compute of the global series (Weighting total of curves (then : functions) which be knowing on different granularities
  - Load-Forecasting using especially data-stream structure.

2. Some preliminary ideas for the exploitation of the data streams for forecasting models

Then, what about load forecasting using data Stream?

- Work of NN Vijayakumar, B Plale, R Ramachandran and X Li (see Vijayakamur and Al, 2006) concerning work of mesoscale weather forecasting (weather Phenomena interesting a zone whose area is about a hundred kilometers) by using the data streams with dynamic filters in order to determine and to analyze some phenomena releases mechanism. .

- Other work on summarizing problems of temporal and space-time data (see Zhang and Al, 2003).

- A methodology, named AWSOM (Adaptive, Hands-off Stream-Mining) making it automatically possible to detect seasonal tendencies or other relevant temporal phenomena within a framework of data stream while using wavelets decompositions (see Papadimitrou and Al, 2003 like Papadimitriou and Al, 2004).

- Approaches concerning the multiple and latent variables regressions by stream incremental analyses (see Teng, 2003).

- And all I try to understand during this Workshop!!…

2. Some preliminary ideas for the exploitation of the data streams for forecasting models

Thank you for your attention !