Data Stream and Load Forecasting : some ideas for a research project at Electricité De France

Alain Dessertaine

EDF R&D OSIRIS
1 avenue du Général De Gaulle
92141 Clamart Cedex
alain.dessertaine@edf.fr

Abstract. In this article, we are going to highlight the interest in using the data resulting from the future 32 millions of communicating meters which will equip all the French customers from now to 2013 in order to build up Court-terms forecasts and Means-terms concerning the EDF electric consumption (total, or by wallets) in an environment of data stream. First of all we will develop our reflections and afterwards we will evoke some tracks of research. These tracks should enable us to approach some modelling and forecasts by aggregation/disintegration of curves, as well as modelling and forecasts on Hilbertian data. Finally we will place side by side these ideas with the Stream-Mining type approaches.

1 Context

1.1 Some words about the potential use of the data Streams at EDF

The volume of data treated and analyzed by EDF is getting increasingly important. The installation of systems of measurement, becoming more and more efficient, will increase consequently this volume. Our aim will be to have a lighting on these data and information delivered in a current way for a better reactivity about some decision-makings. Then, for example, a rise in competence on the use and modelling of structured data stream should allow the calculus and the analysis of monitoring indicators and performances of the power stations of production, in an environment of data stream.

Moreover, the installation planed from now to 2013 of more than 32 millions of communicating meters of all of EDF' consumers should allow a better analysis of the EDF customers' spending patterns. We could also analyze "on line" consumption on more or less incorporated levels and predict it in order to adapt the production and the potential purchases on the electricity markets. Indeed, these meters will be used like sensors in order to measure all the load curves of each customer to very fine temporal granularities (going until the

minute, even the second). Today, only a few tens of thousands of customers[1] have meters to recover curves with measurements every half-hour, even every ten minutes. It is the potential use of these future data which leads us to reflect and to build tracks of work and research that we briefly will present in this article.

1.2 The forecast of electric consumption: Why and how?

The forecast of electric consumption is a very important target for EDF. Thanks to the forecast of long term we can manage the future investments in order to make the park of production evolve adequately, whereas with the forecasts of short term (from 1 hour to ten days) and of medium term (from more than 10 days to a few months) we can adapt the piloting of production equipments and the potential purchase on the financial markets of electricity to the consumption of its customers' wallet. In the last both cases, we forecast total consumptions, or by wallet, by hour or half an hour.

Considering the evolution of the competitive context of the electricity's market, the trade of the forecaster in the energy sector strongly evolved and the consumption forecast became multiform taking into account new needs: from a national forecast to the time half step with a perfect knowledge of past and known explanatory variables (such variables of temperatures and nebulosity), we must now plan individual or portfolios' forecasts with different evolutionary data and an increase in the need for forecast quality taking into account the financial stakes.

Also, we actually are trying to adapt and improve the existing models (nonlinear models, according to temperature, nebulosity and calendar's effects) in order to envisage the most individual consumption, to aggregate those consumptions at best, to build fitted models in those aggregates to improve the forecasts on the total signals.

The use of the data coming from the 32 millions of communicating meters should make it possible to elaborate new tools of decision-making help and forecast, by fitting and using the models and tools resulting from the approaches of treatments of the data stream and Stream-Mining. The installation on the market of system allowing the management and the modelling of such a structure of data should become effective next years. For the decisional components to develop in an immediate future, the commercial software of management of the data stream will not be yet available. Then, the stake for EDF will consist in the thorough knowledge of the principles and the algorithms, today present in the prototypes of research, in order to be precursory in a data processing and modelling in a current way when that is possible.

This document must present the first reflections as well as the research orientations set up by the EDF Research and Development Department to exploit these data at best.

[1] They are the most important customers and several samples of customers. With these samples, we can analyse their uses and build institutional profiles for allocating the total consumption among all the French electricity market distributors.

2 The first reflections and some tracks of research

2.1 Some remarks about the source data

The great potential of these future data will be linked to their space and time characteristics and to their "quasi-continuous" temporal characteristic. This last characteristic will inevitably give a very important lighting respect to the studied phenomenon; indeed, we will be able to check if temporal sampling used today for forecasting (by hour, or half an hour) is the best adapted for a good analysis of the electric consumption of a wallet[2] .

Moreover, it is planed to collect, for some customers having yet accepted it (in a contractual or different way…) data by uses (like the heating, heats it water, kitchen, etc…) and, eventually, some temperatures data inside the residences (even outside…).

Invoicing and the "Règlement des écarts"[3] will require data recoveries, for which rules still have to be established. Will we be able to use exhaustive data on the one hand (given daily, even by hour for all customers), and the results of "space-time" samples on the other hand in order to build historical curve with very short and precise granularities (going until the minute, even the second…)?

Our ignorance concerning the potentially exploitable data will have to direct our choices, step by step, towards some research plans rather than others.

On the contrary it is obvious that we can take into account, nowadays, the space-time and quasi-continuous characteristics of our data in order to conceive our first reflections and research tasks.

2.2 Some remarks on streams and summaries of data

The great volumetry of the data which we have just talked about will imply very strong constraints in term of transmission and storage of the data. So reflections will be useful to define our requirements in term for data to use in order to conceive relevant tools for forecasting.

We can however outline some preliminary remarks and recommendations:

- 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 give a possible definition of this phenomena, we call "foisonnement", hereafter).

---

[2] For example, we notice on the French total curve a peak demand at the end of the afternoon, corresponding to the addition of various phenomena like switching the lights on, using plates and electric furnaces to cook, and starting up some water-heaters etc… The forecast of this phenomena is really important, particularly in winter. However, it "moves" constantly and progressively by the time, because of the sunset variable hours. This phenomenon would be obviously easier to notice and to modelize by using "quasi-continuous" data.

[3] Term given to the "control system" of consumption for all electricity distributors, still mentioned above.

- 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.

Some studies are running on in order to propose space-time sampling strategies of streams to be used[4]; in order to take into account some preliminary constraints to our future treatments, we will refer to these first studies.

So, we know that 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. Let' us imagine as a working hypothesis that we will manage 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. As follows:

- we will be able to collect curves with different, but constant temporal granularities throughout certain periods of each curve (the temporal window could be, for example, daily), these granularities chosen to optimize the total quadratic errors (Chiky, 2007).
- 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.
- we will also be able to collect summaries with functional decompositions (on wavelet basis, for example). Compressions will be able themselves to depend on the complexity of each imported signal and/or each studied moment.

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

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

Some studies on individual curves revealed that to predict the individual consumption was very difficult because of their strongly random and erratic characteristics. As we noticed higher, it would be useful to gather the curves cleverly in order to use the fact that the sum (or the estimate of the sum) of the curves of each group will be foreseeable.

This track is likely a problem of curves classification by tree, including at the same time exogenous variables (space co-ordinates and, if possible, interior and/or exterior temperature), and, especially, an indicator (to be defined) allowing to prune the tree on a adapted level. The target is to improve the predictibility of the global signal after summarizing forecasts on each built class! The variable to explain would not be a simple continuous variable, but a variable made up of functions. Bonds will be undoubtedly created with the works presented in Ramsay and Silverman (2002 and 2005).

[4] Thesis of Raja Chiky (ENST) under the direction of George Hebrail

Moreover, the continual stream would need to update this classification in an incremental way.

2.3.1 A possible formalization of the problem

- Notations :

Let i, an individual on which is measured a curve of consumption of electricity $C_i(t)$.

The overall consumption of the population of this study can be written (we will call the corresponding curve synchronous):

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

with N global strength of our study population. Let us notice, here, that curves C_i will not be inevitably measured with the same step of time, or with the same smoothness. So, that's why we must establish adapted functional smoothings to estimate this sum.

The population of this study can be divided, over one period T given, in G_T sub-groups g of N_g customers. We can define consumption by group in the following way (under-synchronous):

$$C(t) = \sum_{g=1}^{G_T} C_g(t) \tag{2}$$

with :

$$C_g(t) = \sum_{i=1}^{N_g} C_i(t) \tag{3}$$

Let us notice right now that if we work only with one sample of curves of customers, we can calculate instantaneous estimates of these overall consumptions if we know a set of weight w_i: these weights will be, initially, the opposite to inclusions probabilities, directly resulting from the sample design used to build our panel. In that case we have:

$$\widehat{C}(t) = \sum_{i=1}^{n} w_i C_i(t) \tag{4}$$

$$\widehat{C}_g(t) = \sum_{i=1}^{n_g} w_i C_i(t) \tag{5}$$

with n and n_g respectively numbers of customers in the total sample, or the sample on the group g. The formula (5) amounts carrying out the estimate of synchronous on the domain defined by the group g. Of course, some links could be established in order to use all information in our possession to estimate as well as possible $\widehat{C}_g(t)$, particularly in the case of a small domain. For more details, we can refer to Deville and Särndal (1992), Lundström and Särndal (2005) and Rao (2003).

In both cases, we can calculate, even estimate the estimators' sampling errors. It would be convenient to reduce the sampling errors by using all information in our possession, either by building balanced samples, or by carrying out rectifications in order to preserve as well as possible in our samples and our estimations the temporal "representativeness" of our survey.

Some works have been initiated taking into account Hilbertian variables to build Re-weightings calibrations estimators (see Dessertaine, 2006), or balanced sampling (see Dessertaine, 2007).

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

Let us accept the existence of a partition of the customers in G_T groups such as we could model each signal $C_g(t)$ so that the sum of the forecasts calculated over the period T is better than the forecast calculated with an adapted model on the aggregate signal C(t).

On the electric signals of consumption, we can illustrate this phenomena by taking into specific account the local weather variables to model consumption of the thermo-sensitive customers of the same perimeter. Other more specific variables (socio-economic, demographic, contractual etc…) will have obviously to be taken into account on this level!

Another way of describing this principle would be to say that for an individual i , element of a group g, its signal of consumption could be divided in an additive way into two random signals $S_g(t)$ et $\xi_i(t)$:

$$C_i(t) = \lambda_i S_g(t) + \xi_i(t) \tag{6}$$

So, we suppose that the individual signals could be divided in the sum of a common signal to all the individuals of the group g, except for a multiplicative coefficient, *and of a distinctive signal of his.*

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

$$Cov_t(\xi_i(t), \xi_j(t)) = 0 \quad \forall (i,j) \in g \tag{7}$$

with :

$$Cov_i(\xi_i(t), \xi_j(t)) = \int_T \left(\xi_i(t) - \frac{1}{T} \int_0^T \xi_i(k).dk \right) \left(\xi_j(t) - \frac{1}{T} \int_0^T \xi_j(k).dk \right) \tag{8}$$

In this case, we can write:

$$C_g(t) = S_g(t) \sum_{i=1}^{N_g} \lambda_i + o(S_g(t) \sum_{i=1}^{N_g} \lambda_i) \tag{9}$$

This can be a "definition" of the "foisonnement" between curves, generalized with a specific clustering or classification. If we don't work with all of the individuals but with a sampling, we can say that, by using the set of weight w_i :

$$\hat{C}_g(t) = \hat{S}_g(t) \sum_{i=1}^{n_g} w_i \lambda_i + o(\hat{S}_g(t) \sum_{i=1}^{N_g} w_i \lambda_i) \tag{10}$$

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. If we measure the performance by a function Q (RMSE of forecast over one period p given for example), we have:

$$\bar{Q}(\sum_{g=1}^{G_T} \hat{S}_g(t) \sum_{i=1}^{N_g} \lambda_i, t \in p) \leq \bar{Q}(\hat{C}(t), t \in p) \qquad (11)$$

This value depends on the performance of each predictor and, also, on the sampling error carried out on the term $S_g(t)\sum_{i-1}^{N_g}\lambda_i$, if this one is estimated by the value $\hat{S}_g(t)\sum_{i-1}^{n_g} w_i\lambda_i$.

Some ideas for clustering and classification:

The problem is how to build and/or maintain a partition in G_T groups respecting the hypothesis above, for checking (11). Thus, we can elaborate judiciously a clustering or a classification of individual signals available in G_P classes, .
Also, the signal $S_g(t)$ could be approached by the average signal of each individual signal belonging to the class g, at every period T. It "would be enough", within the framework of a hierarchical partition for example, to introduce an indicator to cut of the tree to respect the constraint (11).

But, several questions arise such as :

- How can we make an adapted Compression of each curve? Some interesting tests were carried out by using cleaning and compressions with wavelet in a set of 2300 curves. These tests used principles and algorithms described and suggested in Misiti and Al, (1998) ;
- How can we estimate each distinctive signals $\xi_i(t)$ allowing to check (7)?
- How can we formalize an indicator of cutting allowing to respect (11)?

Other points will have undoubtedly to be arisen on this level (like taking into account the foreseeable characteristics in the classification or clustering algorithm, by the choice and the use of an adapted distance, for example). But, with the using of data stream, we will be interested in the incremental update of these clusters or groups. Some approaches were proposed within the framework of update clusters of purchase sequences contained in batches (see Marascu and Masseglia, 2007). In our case, and to make a parallel with their work, the complete sequences to classify would be present in several batches (new batches coming only to update the panelized historical curves).
Another point to be approached will concern the update of the built panels (either at the time of the rejection or integration of customers). This point will be more simply taken into account within the framework of a supervised classification:

Some ideas for supervised classification:

We will have to discriminate a whole of curves, according to a few number of variables describing the customer or his environment like local curves of temperature, nebulosity or any other history of socio-economic data described on the zones of the customers sampled (data resulting from the censuses of the population to the level of the communes, of the districts etc…). Within a non dynamic framework , preliminary treatments of the curves must be carried out before building classifications by tree (for example). The approaches developed by Ramsey and Silverman (2005) could be tested in this case (particularly considering the hypothesis of different temporal granularities using all over and between each curves to be classified).
Within the "dynamic" framework which will be offered to us by using the data streams, an idea would be to check, on each arrival of new stream, the stability of the partition used.

2.3.2 Some ideas for the forecasting models and approaches incremental:

Once the G_I classes *built* , it will be necessary for us to work out and/or update adapted model to each class. The three following points should help us to build more adapted models than today:

- The data will be very fine temporal and variable by the time granularities
- 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.

Thus, the first point will lead us to study work concerning the forecasts on functional data, including the methods derived from the Hilbertian Auto-Regressif models (see Bosq, 2000). Other very recent work shows a great interest to use some interesting properties of the wavelets for the forecast. On the statistical level, several methods are available: the methods of regression on the wavelets coefficients, the methods resulting from the wavelet spectrum to generalize ARMA models with the processes locally stationary, and, finally, the methods of the nonparametric forecast type but where the similarities are evaluated on the wavelets coefficients and not on the original signal; Some very recent work of Anestis Antoniadis and Theofinadis Sapatinas allowed to successfully develop and test these last approaches, more particularly on data of half an hour electric consumption (see Antoniadis and Al, 2006). The incremental characteristic of this method (comparison of a temporal window which immediately precedes the period by forecast - contained in the last data summarized resulting from last transmitted stream - with the whole of the same windows width in the past for taking into account of their immediate future in the calculation of a forecast with a Kernel approach) could be particularly interesting with a current way approach.

Then, the space characteristic of our data will be very interesting. Indeed, the influence of the temperature or other meteorology characteristics like the propagation of weather phenomena should largely be highlighted and should be taken into account by analysis and models on space data approaches.

If we will work with an aggregation/disintegration of curves treatment by using a panel managed in an environment of data stream, the G_I estimated curves (at the level of the G_I

classes elaborate and maintained specifically) will be known with a sampling errors. Also, the data to be modelled will be a succession of normal distributions, whose variances of the distributions will vary at each time. So, we will have to be used the knowledge of these variances in our modelling works and, also, in the restitution of these variances for the construction of confidence intervals of our forecasts. Currently, the analysis and the modelling of symbolic data, suggested and developed in the years 1990 and 2000 by Edwin Diday (see Billiard and Diday, 2006) seems to be an interesting track to approach in order to model our particular data, comparable with symbolic objects. Some works were developed within a framework of modelling of time series them, and presented by Carlos Maté Jiménez, from the Comillas university of Madrid during the 26th International Symposium of Forecasting. Those works concerned exponential smoothing models on series with histogram values (see Maté Jimenez, 2006). Some studies are running on in order to build autoregressive models on such data.

2.3.3 And forecasting models in an environment of data stream?

It is obvious that an investment around models suggested specifically in an environment of data stream will be set up. Some readings are projected. Currently, 4 approaches hold our attention. Initially, 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 will be able to hold our attention (see Zhang and Al, 2003).

In the same way, 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, could be interesting (see Papadimitrou and Al, 2003 like Papadimitriou and Al, 2004).

Other approaches, rather relating to the framework general of Stream-Mining will be studied (like those developed by Computer Science department of the Stanford university, in Babcock and Al, 2002). The approaches concerning the multiple and latent variables regressions by stream incremental analyses will be naturally studied (see Teng, 2003).

3 Conclusion

In this article, we highlighted the interest in using the data resulting from the future 32 millions of communicating meters in order to build EDF electric consumption short term forecasts (in globality, or by wallets) in a data stream environment. The investment on the search or the development for techniques using this voluminous data in a current way was evoked. First ideas, coming naturally from a logic of treatments and "traditional" models from data and time series have been proposed; those are based on approaches of aggregation/disintegration of curves and on the use of models on Hilbertian or Functional data, but also on adaptations of usual techniques of the surveys to control the phases of sampling and calibration. Some approaches developed within a more specific framework of

Data-Stream and of Stream-Mining must be the subject of the future readings and studies Also, links between the various approaches evoked in this paper will have to be built in order to combine them for the development of methods and adapted models. EDF has still 6 years before the effective use of the recovered data of these meters. It is interesting to take note that a common laboratory between EDF and the French Telecommunication superior national school is set up right now in order to treat of these problems.

Références

Antoniadis A., Paparoditis E. et Sapatinas T. (2006). *A functional wavelet-kernel approach for time series prediction.* Journal of Royal Statistical Society 68 Part 5 pages 837-857

Babcock B., Babu S., Datar M., Motwani R. et Widom J. (2002). *Models and issues in data stream systems.* ACM Symposium on Principles of Database Systems (PODS) 2002

Billard L. et Diday E. (2006). *Symbolic Data Analysis.* Conceptual Statistics and Data mining: Wiley

Bosq D. (2000). *Linear processes in function spaces. Theory and Applications.* Springer Verlag : Lectures Notes in Statistics n° 149.

Chiki R. (2007). *Répartition optimisée des pas d'échantillonnage : Application aux Courbes de charge de consommations électriques*. ECG 2007

Dessertaine A. (2006). *Sondages et séries temporelles : une application pour la prévision de la consommation électrique.* actes des journées Françaises de Statistique 2006

Dessertaine A. (2007 – to be published). *Sampling and Data-Stream: Some ideas to built balanced sampling using auxiliary Hilbertian data.* ISI 2007 (IPM 56 "New methods of sampling")

Deville J.C. et Särndal C.E. (1992). *Calibration estimators in survey sampling.* Journal of the American Statistical Association; 87 : 376-382

Lundström S. et Särndal C-E. et (2005). *Estimation in Surveys with Nonresponse*. Wiley

Marascu A. et Masseglia F. (2007). *Limites d'une approche incrémentale pour la segmentation de séquences dans les flux.* ECG 2007

Maté C., Arroyo J., Muñoz A.et Sarabia A. (2006). *Smoothing methods for histogram-valued time series.* 26th International Symposium on Forecasting. Santander (España). 11-14 Juin 2006

Misiti M., Misiti Y., Oppenheim G. et Poggi J.M. (1998). *Méthodes d'ondelettes en statistique : introduction et exemples.* Journal de la SFDS n° 139

Papadimitriou S., Brockwell A. et Faloutsos C. (2003). *Adaptive, Hands-Off Stream Mining.* International Conference on Very Large Data Bases (VLDB 2003)

Papadimitriou S., Brockwell A. et Faloutsos C. (2004). *Adaptive, Unsupervised Stream Mining* . VLDB Journal 2004

Papadimitriou S., Brockwell A. et Faloutsos C. (2005). *Streaming Pattern Discovery in Multiple Time-Series*. International Conference on Very Large Data Bases (VLDB 2005)

Ramsay J.O. et Silverman B.W. (2005). *Functional Data Analysis*. Springer-Verlag

Ramsay J.O. et Silverman B.W. (2002). *Applied Functional Data Analysis*. Springer-Verlag

Rao J.N.K. (2003). *Small area estimation*. Wiley

Teng W.G., Chen M.S. et Yu P.S.(2003). *A Regression-Based Temporal Pattern Mining Scheme for Data Streams*. International Conference on Very Large Data Bases (VLDB 2003)

Vijayakamur N.N., Plale B., Ramachandran R. et Li X. (2006). *Dynamic Filtering and Mining Triggers in Mesoscale Meteorology Forecasting*. IGARSS, 2006.

Zhang D., Gunopulos D., Tsotras V.J. et Seeger B. (2003). *Temporal and Spatio-Temporal Aggregations over Data Streams using Multiple Time Granularities.* Journal of Information Systems, vol. 28, no. 1-2, pages 61-84