ELECTRIC ENERGY FORECAST FOR THE INDUSTRIAL CONSUMERS

GOUDE Y. *, DAN F.**
*Electricité de France, R&D, Clamart, France
**University of Oradea, Universităţii no.1, Oradea,

Abstract – This paper focuses on short-term electricity load forecasting of industrial consumers. These customers play a major role on the electricity markets and electrical companies have to develop forecasting tools to face this issue. We present here parsimonious and adaptive methods that are able to cope with the common patterns of electricity load curves (intraday and weekly patterns) and can react quickly to abrupt changes in the data. We propose simple statistical models based on linear regression modeling with AR errors and compare them to more advanced statistical models like adaptive Generalized Additive Models. We apply them to two Romanian industrial consumers datasets and we compute post-sample forecasting errors to assess their accuracy for one-day ahead forecast.

Keywords: electricity load forecasting, adaptive forecasts, abrupt changes.

1. INTRODUCTION

Equilibrating electricity generation and demand is a key activity for electricity companies as no perfect technical solution has been found yet to store electricity. To avoid blackouts and financial penalties, electricity providers have to drive their production units in order to satisfy this basic equation and electricity companies have always attached the utmost importance to that issue to manage a wide panel of production units, from nuclear power plants to wind turbines or power dams. The optimization of such a system is a tricky issue and knowing in advance what will future electricity consumption be is essential. That is why electricity forecasts are investigated.

In addition, the constant development of electricity markets in Europe entailed a more and more competitive setting so that electricity providers have to investigate in the field of real time scheduling and adaptive forecasts in particular.

As EDF is the historical provider and because both electricity generation and distribution were a public service in France, EDF used to forecast the whole French consumption. From this long range aggregated time series, highly parametric models were derived that could capture very specific patterns of the French load (see [1] for a description of this model). More precisely, these historical EDF models are based on non-linear regression techniques to modelize the main features of the electricity demand (see [2]):

- Meteorological: temperature, cloud-cover, felt temperature...
- Economical: economic growth of industrial production or more generally GDP's growth plays a major role on the long-term trend of the load.
- Social: different patterns in the electricity load are induced by the occidental way of life, a yearly cycle (public holiday, banking holiday), a weekly cycle (difference between days of the week and week-end) or a daily cycle (night and day, lunch time...).

As this non-linear regression model do not produce perfect forecasts, a SARIMA modelling (see [3]) of the residuals is used for the short term (from intraday forecast to day+1 forecasts) horizon.

Since the opening of the market, EDF portfolio consumption is diverging from the French electricity consumption and EDF has to reconsider its forecasting policy. Two approaches are currently investigated by EDF R&D load forecasting team. One is based on forecasting the EDF portfolio consumption, that will be referred to as the *aggregated approach*. The other possibility is to classify customers into clusters -e.g.: industrial customers, residential customers-, produce one forecast for each cluster and then sum this forecasts to build the EDF portfolio forecast, we will call it the *disaggregated forecasts*.

For the aggregated approach, statistical methods based on adaptive modelling have been experienced by EDF R&D to deal with breaks in the EDF portfolio (due to loss or gains of customer) as historical EDF model faces with an extensive number of parameters and a low ability to adapt to some changes in the data. State-space models and Kalman filter methods were used in that way in [4] and it obtained good results on the French consumption. Semi-parametric approach using Generalized Additive Models (GAM) -see [5, 6]- that can carry out non-linear effects and produce relatively parsimonious models at the same time were also tested on the French load. The results are presented in [7]. A recent extension of GAM methods allows the on-line estimation of the GAM model, i.e. real time update of the model as soon as the data are observed. This adaptive GAM method has been successfully applied to the French load data in [8].

For the disaggregated approach, an important issue is that the modelization has to be thought differently depending on the cluster of customers and obviously, a cluster of residential customers wouldn't have the same properties as a cluster of industrial ones. To our

 ISSN 2067-5534

knowledge, a few work is available in the literature concerning this disaggregated approach. One can find in [9] a method to optimize the clustering to improve the prediction of the aggregated signal; the historical EDF model producing the forecasts. The data considered in this paper are individual customer data, mostly industrials, and it has been shown that the historical EDF model that behaves quite well on the aggregated signal can't fit with all these individual consumptions. That makes sense as industrial processes can be very different from each other and submitted to different hazards.

We propose here to focus on the problem of forecasting individual industrial customers. As EDF individual commercial data are highly confidential, we propose to apply adaptive GAM methods to two Romanian industrial customers provided by the University of Oradea. We argue that non-parametric modelling is particularly useful in this situation as it provides a less restricted framework than parametric models, and thus can cope with different individual features with a few human interventions. Furthermore, the on-line adaptation of GAMs is a way to deal with breaks in the industrial processes.

2. GAM METHODS: A BRIEF INTRODUCTION

We observe data (X_t, Y_t) where Y_t is a real random variable to predict and X_t the p-dimensional t[th] line of the $n \times p$ matrix X of the explanatory variables. The index ***t*** is the time.

A GAM is a Generalized Linear Model with a non-linear part, for example:

$$g(\mu_i) = X_i^* \theta + f_1(x_{1i}) + f_2(x_{2i}, x_{3i})$$

where $\mu_i = E(Y_i)$ and Y ~some exponential family distribution. The f_j are each represented using a linear basis expansion (e.g. a B-spline basis) and X_i^* is a line from the matrix X^* included in X of the variables having a linear relation with Y. In the following, the link function g will be the identity.

Different methods exist in the literature to estimate such non-linear models. Among the most popular ones stands the backfitting algorithm from Hastie and Tibshirani (see [10]) which is implemented in the statistical softwares SAS, S-PLUS and in the *gam* R package. We don't use this algorithm as to our knowledge no on-line update of the model is available for now. Another popular method is the Penalized Iterative Re-Weighted Least Square (P-IRLS) method from Simon N. Wood [6] implemented in the *mgcv* R package. This method is particularly adapted with real time forecasting problems as it allows the on-line update of the GAM model since the 1.7.6 version.

In a schematic form, this consists in the following procedure. First of all, given a -potentially high dimensional- spline basis (ex: B-splines, cubic regression splines...) the algorithm proceeds in an evaluation of the basis over the data X. We denote U the corresponding evaluation which is a $n \times q$ matrix, where q is the dimension of the spline basis. Then, given a vector valued penalization parameter λ -controlling the smoothness of the f_j- the second part of the algorithm consists in solving the following optimization problem to obtain the penalized regression of the data on this spline basis:

$$min_\beta \|Y - U\beta\|^2 + \sum_j \lambda_j \beta^t S_j \beta$$

Where the sum is made over all the non-linear effects in the model, S_j are known matrix corresponding to the spline basis. When λ is fixed this problem is equivalent to a ridge regression problem and can be readily solved. The trickiest point is the choice of the λ parameter. In the *mgcv* package this is done with General Cross Validation methods (GCV). We refer to [5] and [6] for a more in depth description of the method.

3. DATA PRESENTATION

a) The data

For this study, University of Oradea provided us with two industrial consumptions. These are hourly data, from January 1st , 2011 up to March 31st , 2011 with a lot of missing data each month (about 10 days). Furthermore, no explanatory variable were available and we have no information about the industrial processes that drive these consumptions.

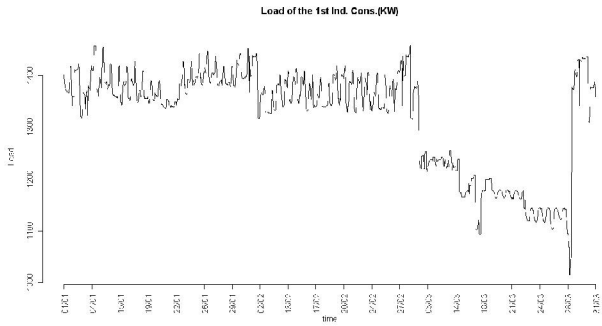


Fig. 1. Electricity load from January to March 2011 for the 1st industrial consumer.

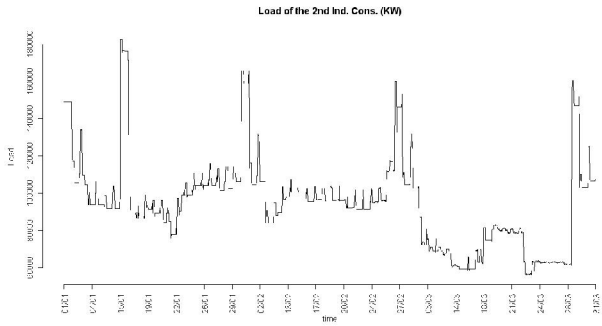


Fig. 2. Electricity load from January to March 2011 for the 2nd industrial consumer.

Figure 1 and figure 2 show the two electricity load curves of these consumers for this period of time after data pre-processing: we have observed outliers –data are

about 1000 times the "normal" load- on the 01/04, 02/04 and 04/04 all day long and 03/31 at noon. These outliers are excluded from the data set.

A preliminary observation is that these two series look very different. The 1st consumer's time series is "smoother" than the 2nd which exhibits some peak days. These peaks would probably be difficult to predict without any exogenous information. Another interesting point is that the two series exhibit breaks at the beginning and at the end of March. As we have no information about the underlying industrial processes and no explanatory variable we have to design adaptive forecasts to cope with these breaks.

b) Basic features of the data

Daily patterns are basic features of electricity load curves (see e.g. [2] or [11]), and that seems to be the case for the 1st consumer and less obvious for the 2nd one as shown in figure 3 where the acf for the two series are plotted.

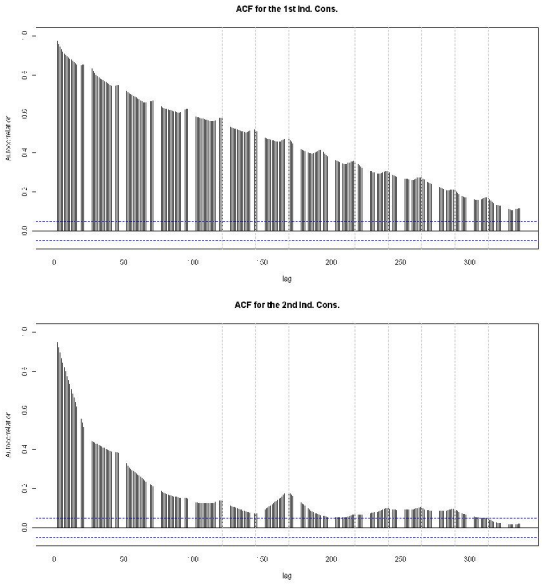


Fig. 3. ACF for the 1st and 2nd industrial consumers

This is confirmed by figure 4 where the box plots of the two signals are plotted relatively to the hour of day. Clearly, the 1st signal has a more significant daily effect that the 2nd one.

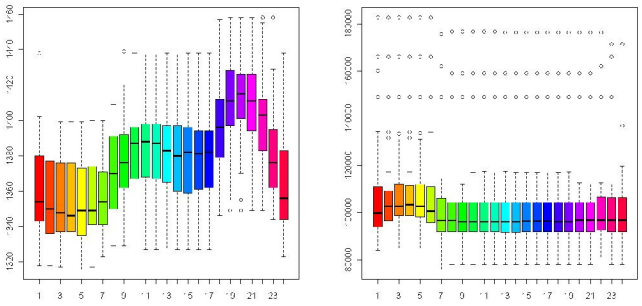


Fig. 4. Box plots of the 1st (left) and 2nd consumptions in function of the hour of the day

We also observe a week days/week-end effect that corresponds to low electricity consumption on weekends. This is shown in figure 5. Once again, the effect is more pronounced for the 1st consumer.

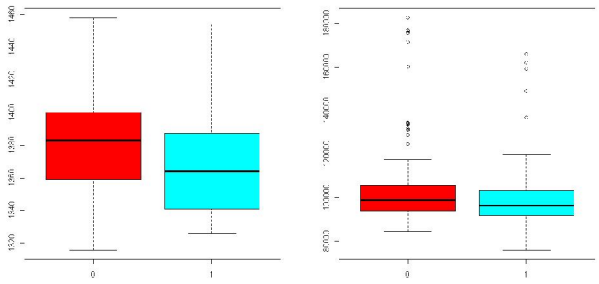


Fig. 5. Box plots of the 1st (left) and 2nd (right) consumptions for week days (in red) and week-ends (in blue).

4. MODEL

In this section we present the first step of our work that consists in model selection and estimation. The philosophy of the work is to derive parsimonious and general models that can adapt to the different features of the two signals presented in section 3 and react quickly to breaks in the data. In a first part, we present a simple model based on basic regression techniques that will be a good benchmark in the following. The estimation of GAM models comes in a second part.

We divide the data set into two parts. The first month (January) is devoted to the estimation of the models. The two following months will be used as a validation set to evaluate the forecasting performances of our models. Note that the estimation set and the validation set are defined in an unbalanced way. This is due to the fact that the estimated models will be updated on-line during the forecasting step, so that the estimation set will grow with time afterwards.

To measure the quality of our models we used two common criteria: the Root Mean Square Errors (RMSE) and the Mean Absolute Percentage Error (MAPE) that are defined as follows. Considering a response variable $y=(y_1,...,y_n)$ and an associated forecast $\hat{y}$ the RMSE and MAPE are:

$$RMSE(y,\hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}$$

$$MAPE(y,\hat{y}) = 100\frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|$$

These criteria are widely used in forecasting literature, the RMSE as quadratic loss is most popular loss in statistics and the MAPE as it expresses accuracy as a percentage and thus is easy to interpret.

a) A regression model

We fit a regression model that expresses the basic data features exposed in section 3. This model is described in the following equation:

ISSN 2067-5534 © 2011 JSE

$$y_t = \sum_{q=1}^{2}\sum_{j=1}^{24} a_j^q I\{h_i = j\} I\{d_i = q\} + \varepsilon_t$$

where h_i and d_i are respectively the hour and the day type observation i and ε_t is supposed to be an AR model: $\varepsilon_t = b\varepsilon_{t-24} + u_t$ where u_t is an i.i.d. white noise. We consider two day types: week days and week ends. This model simply fit two different hourly patterns for this two day types, plus a lag effect of the forecasting error observed a day before. The choice of this autoregressive structure is made according to the daily seasonality observed in figure 3 and also to allow this model to produce day ahead forecasts –intraday correlations surely exists and should be exploited for intraday forecasts-.

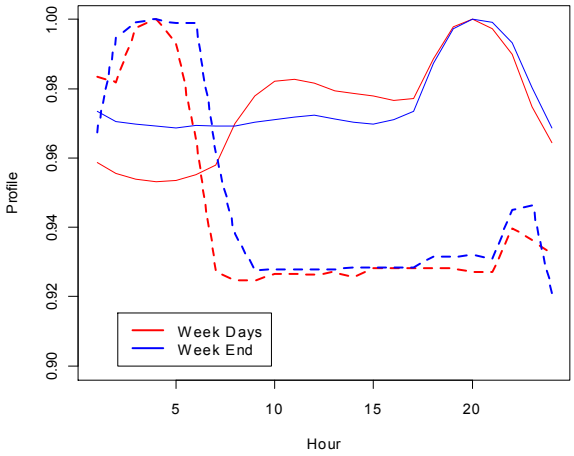
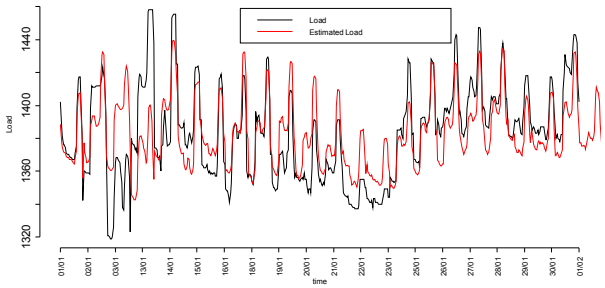


Fig. 6. Scaled (divided by their maximum values) daily patterns for week days (in red) and week end estimated on the 1st consumer (solid line) and the 2nd one (dashed line).

We proceed to the estimation of this model in two steps. First we estimate the fix patterns, and then we fit the AR model on the residuals of this estimation. Figure 6 shows the scaled estimates of these daily patterns for the two signals. The daily patterns are very different on the two signals and the distinction between weekdays and weekends is more pronounced for the 1st consumer.



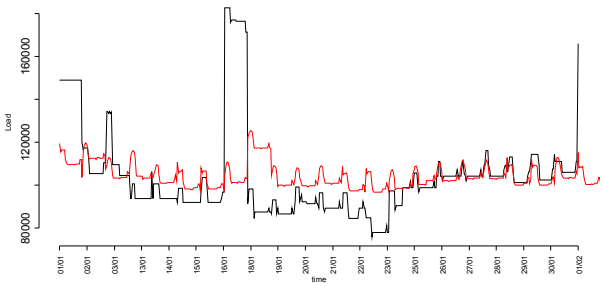


Fig. 7. Load (in black) and estimated load (in red) over the estimation set (January 2011).

The estimated load obtained from this model is presented on figure 7. As expected, the best fitting is obtained for the 1st consumer and the estimation error goes from 1.1% for the 1st consumer to 11.7% for the 2nd consumer. Notice that for the 2nd consumer the estimation error is higher after AR correction than before. That is due to the peak day observed on January the 6th. At that time the estimation error is high and this error is regrettably propagated over the next days.

Table 1. Estimation results on the two signals RMSE and MAPE for the linear regression model.

	Daily patterns	Daily patterns+AR model
1st consumer	22-1.3%	20-1.1%
2nd consumer	20596-11%	21088-11.7%

The auto-correlations are also checked and confirm the proposed AR model.

b) GAM fitting

The estimation and selection of the GAM consist in stepwise procedure where we successively test different modifications. As selection criteria we used the adjusted R-square and Student tests of the different linear and non-linear effects, we refer to [6] for a presentation of this indicators. This selection procedure is done for the 1st consumer, as it is the most "stable" consumer. We finally obtained the following model:

$$y_t = \sum_{q=1}^{2}\sum_{j=1}^{24} a_j^q I\{h_i = j\} I\{d_i = q\} + s(y_{t-24}) + s(t) + \varepsilon_t$$

Where the notations are the same than for the regression model y_{t-24} is the 24-hour lag of y_t and ε_t is an iid white noise. The main differences with the regression model presented in a) are:

- a non-linear effect $s(y_{t-24})$ for the 24 hour auto-correlation
- a non-linear trend $s(t)$

The non-linear effect plays an important role as it allows the off-line estimation of smooth changes in the data. This is presented on the figure 8.

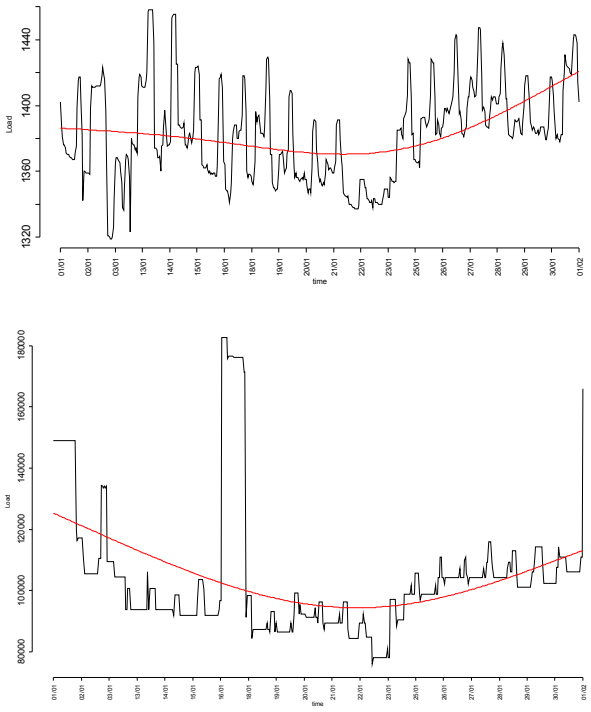


Fig. 8. Load of the industrial customers and the estimated non-linear trend.

The estimation errors we obtained are summarized in the table 2. According to this statistics one can roughly assess that the estimation fitting is better for GAM than for the linear regression model, on the two signals. Actually, as we introduce non-linear effect in GAM, the degrees of freedom of these models are greater than for the linear model. More precisely, the *mgcv* package produce an estimated degree of freedom and our GAMs have nearly 5 degrees of freedom more than the linear model and this improvement can be due to overfitting. We'll see in the next section that it is not the case.

Table 2. Estimation results on the two signals - RMSE and MAPE for GAM.

	GAM-RMSE	GAM-MAPE
1st consumer	16	0.90%
2nd consumer	18733	9.60%

5. FORECASTING

This last section is devoted to the forecast of the two industrial consumption signals. As previously described we forecast this time series over the two last months of the data (February and march 2011) and we allow the on-line update of our two models. Each 24 hours, the parameters of the regression model and the GAM are updated before a new prediction is made. The forecasts we obtain over the 2 months are shown in figure 9. The forecasting results are presented in the table 3.

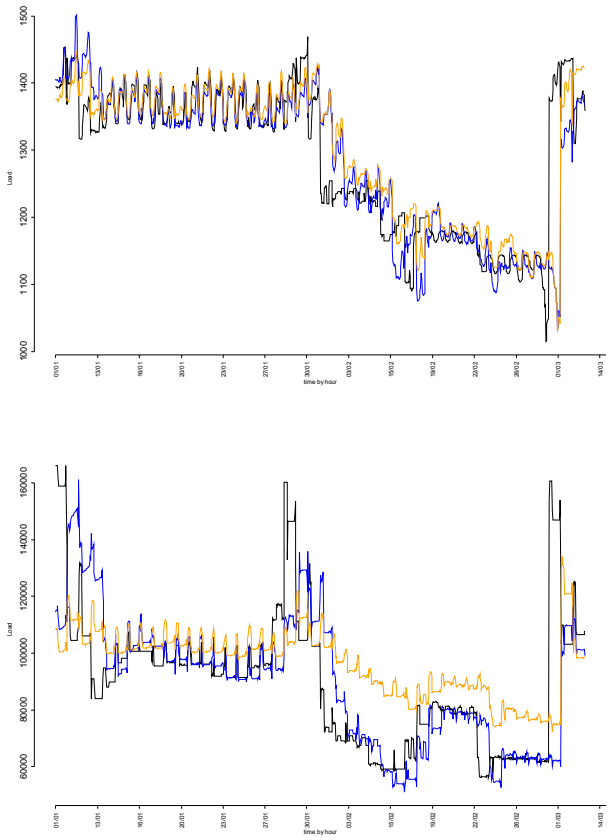


Fig. 9. Load and 1 day ahead predicted load for adaptive reg.-AR model (in orange) and adaptive GAM (in blue).

As expected, the forecasts results are better on the first signal, with a 2.7% MAPE. This is a quite good performance as we forecast without any exogenous information. In addition, bad forecasts correspond to the breaks at the beginning and at the end of March as the models adapt progressively to these changes and exhibit a kind of inertia to the breaks. Anyway, the two models realize slightly the same performances on the 1st consumer signal. An interesting point is that the GAM is really better than the adaptive linear regression model on the 2nd signal. This is quite obvious after the second pick day that occurs at the beginning of March. The linear regression model adapts really slowly to this break in comparison to adaptive GAM.

Table 2. Forecasting results on the two signals RMSE and MAPE for the two adaptive models.

	Adaptive Reg.-AR	Adaptive GAM
1st consumer	61-2.7%	63-2.7%
2nd consumer	21837-18.9%	20804-11.6%

Actually, the use of exogenous information should be very useful, especially for the 2nd signal where the peak days have a harmful effect. We proceed to a last run of forecasts on the signal, excluding the peak days and we obtain the forecasts presented on the figure 10.

ISSN 2067-5534 © 2011 JSE

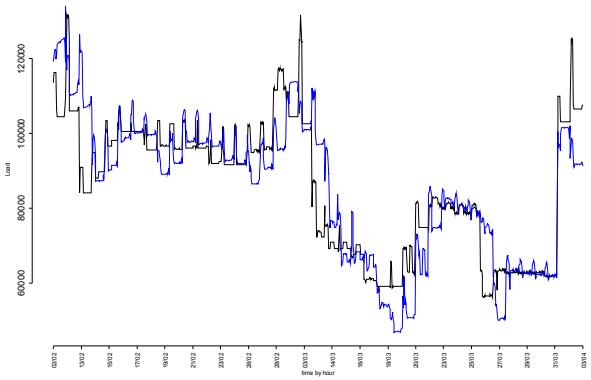


Figure 10 Load and 1 day ahead predicted load for adaptive GAM (in blue).

These forecasts achieve an 8.3% MAPE and that confirms the importance of peak days in the final performance of our forecasts, highlighting the importance of exogenous information on industrial processes that drive the load curve.

5. CONCLUSION

In this paper, we address the problem of forecasting industrial customers demand in a disaggregated way. Thus, we had to work on short time series of only 3 months whereas modelling electricity load often requires several years of data. Furthermore, the two individual consumption signals were shown to have different properties, as well as abrupt changes. To solve this tricky problem we proposed two adaptive methods: basic regression methods and AR models, as well as GAM models. The adaptation to the data is made on-line, in a real time update fashion. We showed that adaptive GAM can be a good solution as it reacts to changes and can adapt to different individual signal thanks to its non-parametric structure.

A future work should be to test these methods on other industrial consumptions. Another improvement should be done to produce intraday forecasts, as we believe that abrupt changes can be easier to detect at this time scale.

REFERENCES

[1]. Bruhns, A., Deurveilher, G. And Roy J.S. -A non linear regression model for mid-term load forecasting and improvements in seasonality, in Proc. of the 15th Power Systems Computation Conf., Liege Belgium, 2005.

[2]. Bunn, D.W. and Farmer, E.D. -Comparative Models for Electrical Load Forecasting, New York: Wiley, 1985.

[3]. Box, G. and Jenkins, G. -Time series analysis: Forecasting and control, San Francisco: Holden-Day 1970.

[4]. Dordonnat, V., Koopman, S. J., Ooms, M., Dessertaine, A., and Collet, J. -An hourly periodic state space model for modeling French national electricity load, International Journal of Forecasting, Vol. 24, pp. 566–587, 2008.

[5]. Wood, S.N. -Modelling and smoothing parameter estimation with multiple quadratic penalties, Journal of the Royal Statistical Society (B) 62(2): 413-428, 2000.

[6]. Wood,S.N. -Generalized Additive Models: An Introduction with R, London: Chapman & Hall, 2006.

[7]. Pierrot, A. and Goude, Y. -Short-Term Electricity Load Forecasting With Generalized Additive Models, to appear in the 16th ISAP-power proceedings, 2011.

[8]. Wood, S., Goude, Y. and Shaw, S. -Generalized additive models for large datasets, submitted to JASA, 2011.

[9]. Misiti, M., Misiti, Y., Oppenheim, G. and Poggi,J.M. -Optimized Clusters for Disaggregated Electricity Load Forecasting, REVSTAT–Statistical Journal, 8(2), 105–124, 2010.

[10]. Hastie, T. and Tibshirani,R. -Generalized Additive Models. London: Chapman & Hall, 1990.

[11]. Taylor, J., de Menezes,L. M. and McSharry, P. E. -A comparison of univariate methods for fore-casting electricity demand up to a day ahead, International Journal of Forecasting, vol. 22, pp.1-16, 2006.