Short-Term Electricity Load Forecasting With Generalized Additive Models

Conference Paper · January 2011

2 authors:

Amandine Pierrot
Électricité de France (EDF)

Yannig Goude
Électricité de France (EDF)

Short-Term Electricity Load Forecasting With Generalized Additive Models

Amandine Pierrot, *Electricité De France*, and Yannig Goude, *Electricité De France*

***Abstract*--Because of the French electricity market deregulation, Electricité de France has to experiment new load forecasting models, more adaptive than the operational ones. A statistical framework like Generalized Additive Models allows us to integrate both a regressive part with explanatory variables and an autoregressive part with lagged loads. The French electricity demand being strongly related to the current instant, we consider twenty-four daily time-series and fit one model for each hour. The selected variables are one-day-lagged loads, weather and calendar variables, and a global trend. Thanks to a cyclic spline fitted on the position of the current day during the year, we can model the summer break (a large decrease in the demand due to summer holiday). We compute the Root Mean Square Errors over one post-sample year to assess its accuracy for one-day ahead forecast. Our model, which is fitted over five years, can compete with the operational one.**

***Index Terms*--load modeling, power demand forecasting, regression, smoothing methods, spline.**

I. Introduction

Modeling and forecasting the electricity load at short-term and middle-term horizons is a key activity for electrical companies. The need to maintain the equilibrium between the electricity supply and demand at any time is essential to avoid power systems injuries and blackouts that generate financial penalties or more important drawbacks. The French electrical load company Electricité De France (EDF) has always attached the utmost importance to that issue which stands for a central point in power system scheduling, in the management of a large and varied panel of production unit that contains nuclear and thermal power plants, power dams and wind turbines, and for which the optimization of the production costs is essential.

The advent of the wholesale electricity market in Europe and in France (since 2003 for industry, since July 2007 for every customer) has brought renewed focus on load forecasting for different reasons. As the electricity can be bought on a competitive market, EDF needs to include electricity price forecasting in the decision process that results to the power system schedule to maximize its margin. Actually, a marginal improvement of the load forecasts can lead to important benefits, especially during peak load periods or hard to predict situations when the electricity prices can attain very high levels. Another reason for this renewal interest is that the French electricity demand is highly related to technological evolutions in the fields of new technologies (computers, flat screens, cellular phones, heat pumps...). That tends to disturb the very accurate forecasts produced by the historical EDF model based on twenty years of developments and operational feedbacks.

Statisticians and forecasters have produced a large amount of work about short-term (one-day ahead) load forecasting so that proposing an exhaustive survey is a hard task. We only highlight here some popular statistical approaches and results about electricity load data. Many covariates were shown to have a significant effect on electricity consumption, see e.g. [9] and [10]. These covariates can be summarized into three groups: **1. Social**: modern occidental societies exhibit different common features that induce different patterns in the electricity load, for example: 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...). **2. Natural**: temperature, cloud cover, wind speed and day length are important drivers of electricity consumption due to electrical heating and lightning. **3. Economical**: economic growth of industrial production plays a major role on the long term trend of the load.

To deal with seasonal patterns, ARIMA models or exponential smoothing have been proposed. It can be referred to [1] for a survey about univariate methods. State space models came as an extension of ARIMA modeling and including meteorological covariates as in [2]. The most important drawback of these models is their relative difficulty to estimate the non-linear effects of meteorological variables. That can be done with multivariate regression techniques, also popular and widely used in industry, see e.g. the extensive regression model by hour of the day built in [3] and [4], and the nonlinear regression model developed by EDF R&D in [15]. These approaches classically face with an extensive number of parameters and a low ability to adapt to some changes in the data. As an alternative, Bayesian methods were proposed in [6] to face with the intraday patterns of the load, assuming a specific but sparse correlation structure of the residuals. Another popular and powerful approach is artificial neural networks (ANNs) -see [7] and [8]-. ANNs can deal with non-linear effects as well as seasonal patterns, but for an

A. Pierrot is with the Research and Development Division, Electricité De France, 1 av du Général de Gaulle, 92141 Clamart Cedex, FRANCE (e-mail: amandine.pierrot@edf.fr).

Y. Goude is with the Research and Development Division, Electricité De France, 1 av du Général de Gaulle, 92141 Clamart Cedex, FRANCE (e-mail: yannig.goude@edf.fr).

industrial use one can regret their lack of interpretability.

We propose here a semi-parametric approach using **G**eneralized **A**dditive **M**odels -see [11]-[12] for a presentation- that can carry out non-linear effects and produce relatively parsimonious and interpretable models at the same time. Semi-parametric additive models have already been applied to long and short term load forecasting for Australian regions -see [13] and [14]-, and exhibited good performances. We apply GAM methods on the French load consumption data, which has never been done yet, using the *mgcv* R package presented in [12]. We focus first on modelizing the different effects that classically drive the French load consumption. Then, we present the sequential procedure that lead to the final model and compare its forecast to the operational one. We show that this new model is competitive with the EDF operational model.

II. Technical Work Preparation

A. The Data

We fit our model on French load hourly data, from September 2000 to August 2005.

As explanatory variables, we use meteorological hourly data: the temperature, the cloud cover and the wind speed. The cloud cover is from 0 for no cloud to 8 for a very cloudy weather; the wind speed from 0 to 10 m/s usually. These three national variables are computed over 26 stations in France, as a weighted mean. The wind variable is a new variable to be tested, that we get thanks to the progresses in meteorology.

The many different effects in the French electricity demand make its modeling and forecasting a difficult task. That includes a yearly (Fig. 1), a weekly and a daily seasonality (Fig. 2), a growing trend (Fig. 1), a dependence on the temperature and big differences between the summer and the winter due to the use of electrical heating (Fig. 2).

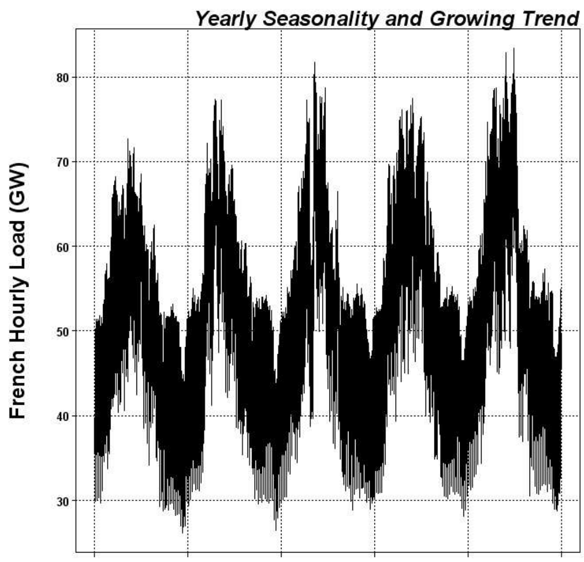

Fig. 1. The French load from 2000 (September 1) to 2005 (August 31). Note the growing trend and the yearly pattern due to economical cycle, heating in winter and summer holiday.

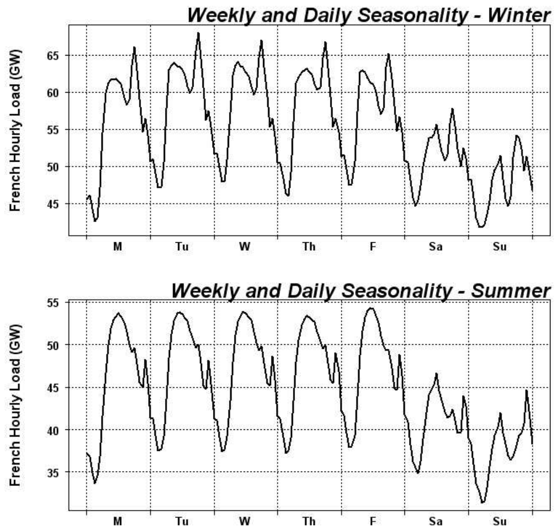

Fig. 2. The French load from Monday to Sunday, in November and in June. Note the weekly pattern: week days are very similar, while Saturday and Sunday are different in both level and form. Note again the differences in the load depending on the season: the day peak moves from around 7 p.m. in winter to 1 p.m. in summer.

Moreover, there are some important decreases in the load during Christmas holiday, summer holiday, banking holiday and special tariff days.

B. The GAM Statistical Framework

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

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

where $\mu_i = \mathrm{E}(\mathrm{Y}_i)$ and Y ~some exponential family distribution. Y is a response variable, g a specified link function, X^* the matrix of the covariates having a linear relation with Y and the f_j are smooth functions of the covariates x_j. The relationship between the variable to explain and the explanatory variables is more flexible because possibly non-linear.

There are several methods to estimate GAM, one of the most famous ones being the backfitting algorithm from Hastie and Tibshirani -see [11]-, which is implemented in the statistical software S-PLUS, SAS and in the *gam* R package. We choose to rather use the Penalized Iterative Re-Weighted Least Square (P-IRLS) method from Simon N. Wood -see [12]-, especially because it includes the estimation of the smoothing degree for the f_j .

In the P-IRLS method, which is implemented in the *mgcv* R package, the f_j are represented using regression splines. Given such bases, a GAM can be estimated as a GLM. To avoid overfitting, *mgcv* controls the smoothness for each term through a set of penalties applied to the likelihood of the GLM. The GAM is fitted by iterative minimization in β, given λ_j, of the problem

$$-l(\beta)+\sum_j \lambda_j \beta^T S_j \beta \qquad (2)$$

where β contains θ and all the β_j, coordinates of f_j in its spline basis, S_j are matrices of known coefficients such that the second term in (2) penalizes models with "wiggly" f_j. l is the log-likelihood for β. The λ_j are smoothing parameters that control the trade-off between fit and smoothness, and can be selected by minimization of the Generalized Cross Validation (GCV) score.

C. Estimation: How Do We Proceed

With the *mgcv* R package, we are not restricted to models including only splines of one predictor, and we also can fit variable coefficient models. For example:

$$g(\mu_i) = \mathrm{X}_i^* \theta + f_1(x_{1i})x_{2i} + f_2(x_{3i}, x_{4i})x_{5i} \qquad (3)$$

Furthermore, the *mgcv* R package gives some indicators of the model fitting quality: adjusted R², percentage deviance explained and GCV score.

The French electricity load being strongly related to the current instant, we consider 24 daily time-series and fit a model for each hour. We then fit 24 models. Table I provides the different effects impacting the French demand, and the corresponding potential variables to model them.

The daily seasonality is modeled fitting a model per hour. The weekly seasonality is modeled using a categorical variable depending on the type of the day. For example, if the different levels are "week day" and "week-end day", two levels of the load are estimated: one level for weeks, one level for week-ends. Of course, this categorical "type of the day" variable may have different levels. This is part of our search for the best model.

TABLE I

EFFECT	VARIABLE TO MODEL IT
economic growth	trend
yearly seasonality	position of the day during the year
weekly seasonality	type of the day
dependence on the weather	temperature of the current hour, lagged temperatures
	cloud cover
	wind speed

Our 24 hourly models are fitted over five years, from September 2000 to August 2005. Christmas and summer breaks, special tariff days and bank holiday are invalid data.

When fitting, we make sure that the residuals of our model are good: we check that the normality hypothesis is not violated, there is no pattern left in the variance or in the mean, the residuals are well-scattered around 0, and there are no auto-correlations remaining. We choose one model over another according to the residuals and the *mgcv* measures of the fitting quality. The GCV score is useful in comparing models with different variables included (the smaller the GCV score is, the better the model is).

We start fitting the noon model. We run some simulation tests and notice that a GAM can deal with some auto-regressive variables. Therefore, even if the theory still does not exist, we include lagged loads into our model, in order to improve it and to reduce potential auto-correlations in the residuals.

We use thin plate regression spline bases: they can smooth any number of covariates, avoid putting "knots" and have some optimality properties. The choice of basis dimensions amounts to set the maximal possible degrees of freedom allowed for each model term. The number of effective degrees of freedom is estimated from data by GCV.

D. Estimation: The Best Model Selection

Our first model, M1, is the following one:

$$\begin{aligned} L_t = {} & f_1(L_{t-24}) + f_2(L_{t-168}) + f_3(T_t) \\ & + f_4(mean(T_{t-24}, T_{t-48})) \\ & + f_5(cc) + f_6(posan) + C + \varepsilon_t \end{aligned} \qquad (4)$$

where L_t, L_{t-24} and L_{t-168} are respectively the t-instant load to be forecasted, the one-day-lagged load and the one-week-lagged load. T_t, T_{t-24} and T_{t-48} are the t-instant temperature, the one-day-lagged temperature and the two-day-lagged temperature. *cc* is the cloud cover, *posan* the position of the day through the year (from September to August), *C* is an intercept and ε_i the residual error.

We introduce a new categorical variable in order to model the weekly seasonality: that is a "type of the day" variable with two levels, a "week" level and a "week-end" level. We compare the results of M1 and M2 (5) in Table II.

$$\mathrm{M2} = \mathrm{M1} + \text{week/week-end variable} \qquad (5)$$

TABLE II

	R²	Explained Deviance	GCV Score
M1	0,911	91,40%	5,69*10^6
M2	0,979	98%	**1,38*10^6**

The introduction of a variable to estimate a different effect for weeks and week-ends really improve the GCV score of the model.

Fig. 3 represents four effects of the model M2: the temperature effect, the lagged temperature effect, the yearly seasonality and the cloud cover effect. The French load demand strongly depends on the outside temperature. The colder the temperature is, the bigger the needs in electricity are, for heating. There is also a slight cooling effect, when temperature is over 20°C. The lagged temperatures are significant because of the temperature reaction inertia. Finally, the power demand is growing when the weather is very cloudy (for a cloud cover bigger than 6). That can be explained by both a cold feeling amplified by clouds and bigger needs in lighting.

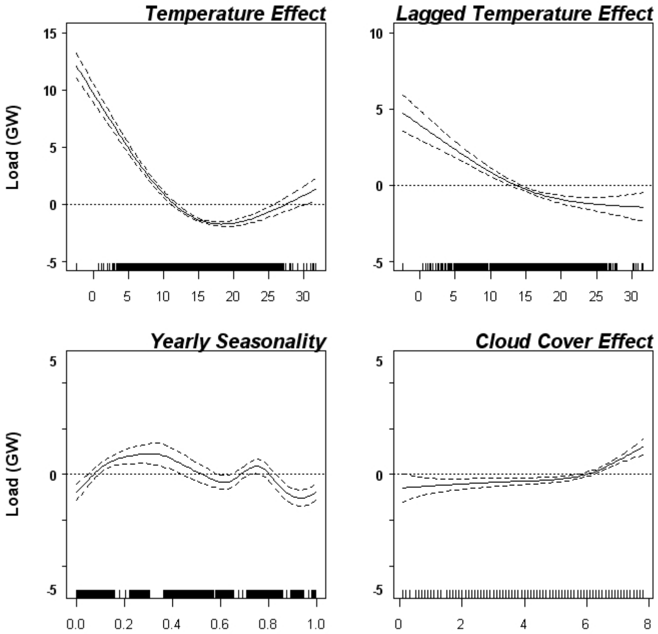

Fig. 3. The French load as a function of the temperature, of lagged temperatures, of the position of the day through the year and of the cloud cover. Note that *mgcv* fits centered functions. That is because of the identifiability constraint.

Fig. 4 shows the model M2 residuals:

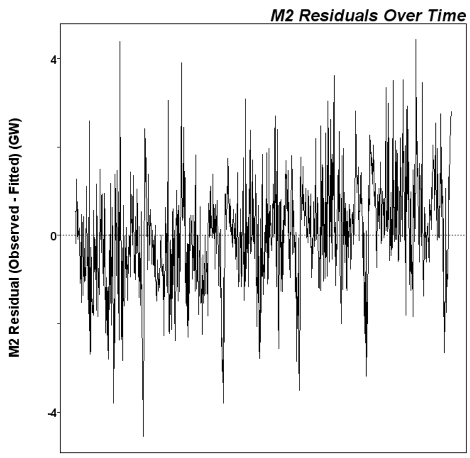

Fig. 4. The estimation error over the five years of fitting, on valid data only. Note the growth.

The M2 estimation error confirms the need for a global trend (6).

$$\mathrm{M3} = \mathrm{M2} + \text{global linear trend} \tag{6}$$

The introduction of the trend to model the economic growth improves fitting *mgcv* indicators (Table III).

TABLE III

	Explained Deviance	GCV Score	New Variable
M1	91,40%	5,69*10^6	
M2	98%	1,38*10^6	week/week-end
M3	98,70%	**8,83*10^5**	trend

The new residuals are on Fig. 5:

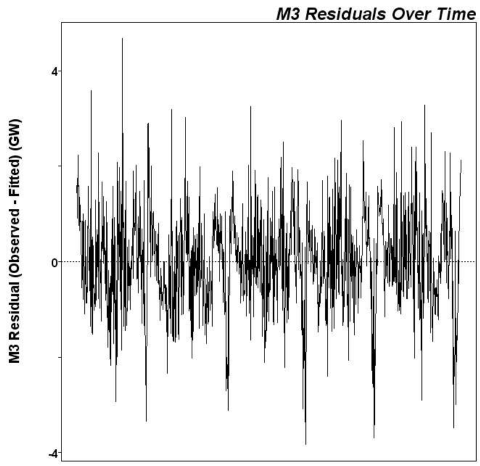

Fig. 5. The now centered estimation error over the five years of fitting.

If we refine the “week/week-end” variable to a “day of the week” variable (a categorical variable with seven levels, one for each day of the week, from Monday to Sunday):

$$\mathrm{M4} = \mathrm{M3} + \text{day of the week} \tag{7}$$

TABLE IV

	Explained Deviance	GCV Score	New Variable
M1	91,40%	5,69*10^6	
M2	98%	1,38*10^6	week/week-end
M3	98,70%	8,83*10^5	trend
M4	99,40%	**3,94*10^5**	day of the week

Once again, the GCV score is clearly improved. If we check the hypotheses on errors (Fig. 6), our model residuals are really good. It seems to be no auto-correlations left.

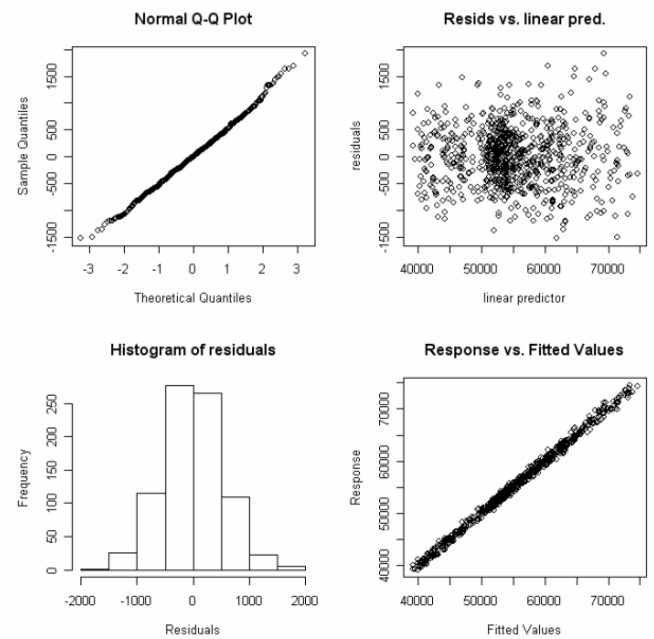

Fig. 6. Four residual plots giving some diagnostic information about the fitting procedure and results.

After a more thorough study of the residuals of the 24 hourly models, we conclude that each function of the one-day-lagged load has to depend on the day-of-the-week variable. Moreover, the level difference between the load to be forecasted and the one-week-lagged load is too variable, and the function of the position of the day during the year already models a middle-term level of the load. That's why we drop the one-week-lagged load.

The final choices that improve our model are the inclusion of the wind speed, of the minimal ($\underline{T_{23}}$) and the maximal ($\overline{T_{23}}$) temperatures for the past 23 hours, and the use of both one-day-lagged and two-day-lagged temperatures instead of their mean.

Fig. 7 presents the model residuals on 4 summer breaks (2001, 2002, 2003 and 2004). We notice that the *posan* variable can deal with the summer break in the load level.

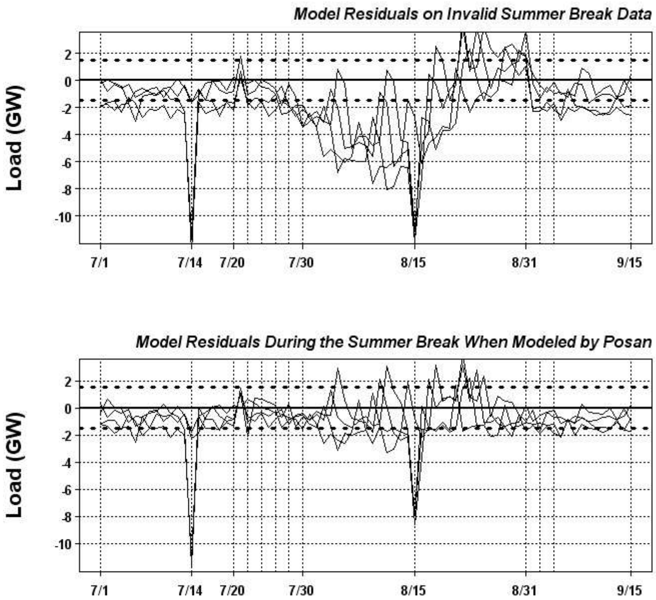

Fig. 7. Model residuals from July to September, before and after the modeling of the summer break by the *posan* variable. Note the overestimations of the load on 7/14 and 8/15, which are banking holiday in France.

This is a very important result. Now we can valid the summer break data and fit a cyclic smooth of the *posan* variable to model it, directly in our final GAM (8), which is not possible in EDF operational forecasting models. The estimated effect of the summer break is shown on Fig. 8.

$$\begin{aligned} L_t = {} & f_1(L_{t-24}).1_{DayOfWeek} + f_2(T_t, wind) \\ & + f_3(T_{t-24}) + f_4(T_{t-48}) + f_5(\underline{T_{23}}) + f_6(\overline{T_{23}}) \\ & + f_7(cc) + f_8(posan).1_{Week/Sat/Sun} \\ & + \lambda.DayOfWeek + \beta.trend + C \end{aligned} \tag{8}$$

Our model consists of a short-term part, a meteorological part and a calendar one. The yearly seasonality is different depending on the type of the day (week day, Saturday, or Sunday), especially during the summer break (Fig. 8). The model equation (8) is the same for the 24 models. Indeed, GAM with *mgcv* R package is a statistical framework adaptive enough, and if in the night models, for example, the cloud cover effect is not significant, the estimated function naturally fits.

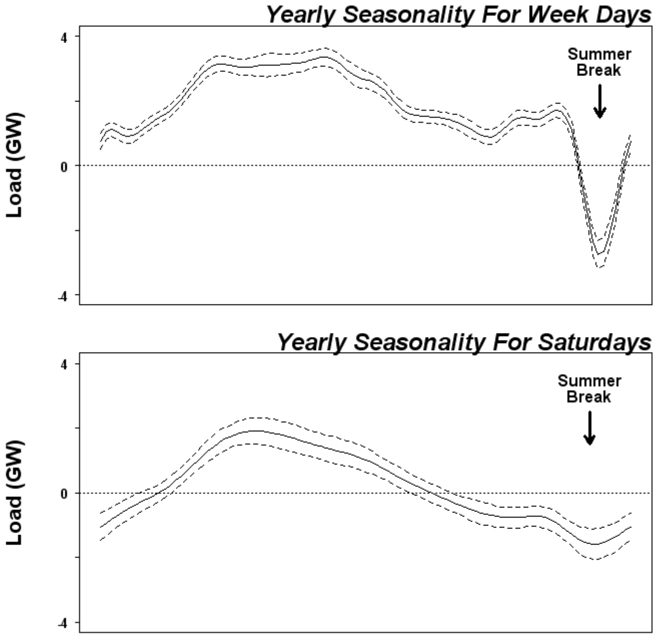

Fig. 8. Yearly seasonality (from September to August): French load as a function of *posan*. Note the difference on the summer break depending on the type of the day, and the growth in non climatic winter load, due to lighting.

E. Forecasting Results

We compute Root Mean Square Errors (RMSE) over one post-sample year (from September 2005 to August 2006) to assess our model's accuracy for one-day ahead forecast. Christmas break, special tariff days and bank holiday are invalid data. The global RMSE of the operational EDF model is about **800 MW** (**1.16%** for the global MAPE), the summer break being excluded. The global RMSE of our GAM (8) is about **600 MW** (**0.90%** for the global MAPE), the summer break being included.

F. Prospects

We keep working on Generalized Additive Models for modeling not only France load but also EDF load and smaller EDF portfolio segment loads. We are especially interested in taking into account correlations between hours, fitting a GAM over the day instead of 24 hourly GAM. That would be a first step in developing a model for intraday forecasting, with lagged loads of the previous 24 hours (e.g. one-hour-lagged load).

We are doing some research about ways to estimate "on-line" Generalized Additive Models, in a data stream context. This kind of model could be very useful if a break in the load level occurs, because of an economic crisis, or customer losses. Moreover, we are planning to test GAM with macro-economic variables, still in order to model the breaks, or the growths, that do not depend on meteorological conditions.

III. References

Periodicals:

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

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

[3] R. Ramanathan, R. Engle, C.W.J. Granger, F. Vahid-Araghi, and C. Brace. "Short-run forecasts of electricity loads and peaks", *International Journal of Forecasting*, vol. 13, pp. 161-174, 1997.

[4] L.J. Soares, and M.C. Medeiros, "Modeling and forecasting short-term electricity load: A comparison of methods with an application to Brazilian data", *International Journal of Forecasting*, vol. 24, no. 4, pp 630-644, 2008.

[5] B. Kermanshahi, "Recurrent neural network for forecasting next 10 years loads of nine Japanese utilities," *Neurocomputing*, vol. 23, pp. 125–133, 1998.

[6] R. Cottet, and M. Smith, "Bayesian modeling and forecasting of intraday electricity load", *Journal of the American Statistical Association*, vol. 98, pp. 839–849, 2003.

[7] H. Hippert, C. Pedreira, and R. Souza, "Neural networks for short-term load forecasting: A review and evaluation," *IEEE Trans. Power Syst.*, vol. 16, no. 1, pp. 44 – 55, 2001.

[8] Z. Yun, Z. Quan, S. Caixin, L. Shaolan, L. Yuming, and S. Yang, "RBF neural network and ANFIS-based short-term load forecasting approach in real-time price environment," *IEEE Trans. Power Syst.*, vol. 23, no. 3, pp. 853 – 858, 2008.

Books:

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

[10] R. Weron, *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach*, Wiley, Chichester, 2006.

[11] T. Hastie, R. Tibshirani, *Generalized Additive Models*. London: Chapman & Hall, 1990.

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

Technical Reports:

[13] Rob J. Hyndman (2007), Extended models for long-term peak half-hourly electricity demand for South Australia, Report for Electricity Supply Industry Planning Council (SA). Monash University Business and Economic Forecasting Unit Available: http://monash.academia.edu/RobHyndman/Papers/287125/Modelling_Long-Term_Peak_Half-Hourly_Electricity_Demand_for_South_Australia

[14] Shu Fan and Rob J Hyndman, "Short-term load forecasting based on a semi-parametric additive model", Available: http://robjhyndman.com/papers/2010STLF-FinalR1.pdf.

Papers from Conference Proceedings (Published):

[15] A. Bruhns, G. Deurveilher and J. S. Roy, "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.