# A NON-LINEAR REGRESSION MODEL FOR MID-TERM LOAD FORECASTING AND IMPROVEMENTS IN SEASONALITY

Alexander Bruhns, Gilles Deurveilher, Jean-Sébastien Roy
Electricité de France
Clamart, France
alexander.bruhns@edf.fr, gilles.deurveilher@edf.fr, jean-sebastien.roy@edf.fr

**Abstract - The aim of this paper is to describe the latest load forecasting model used at EDF for mid-term load forecasting with a particular focus and experiments on seasonality. Its precise description is given along with a discussion on design choices and differences from previous models. This paper provides a short literature survey of methods for modeling yearly electrical seasonality in mid-term models. We present some approaches to improve the modeling of seasonality, using basis function methods such as cubic splines or radial functions and a nonparametric local regression (LOESS). We also test an alternative method to deal with the modification of the daily load shape throughout the year, by introducing two Fourier series: one with dependency on the hour and one with dependency on the day-type. This last model proves to be the best approach, both in accuracy and parsimony, but requires great care in the day-type typology.**

***Keywords - Mid-term load forecasting, Seasonality, Fourier series, Cubic splines, LOESS, Radial functions***

## 1 INTRODUCTION

Since the advent of electricity markets, the need for accurate electrical load forecasting has increased. Operators still run short-term forecasts (five minutes to one week ahead) to ensure system stability, mid-term forecasts (one week to a year ahead) for generation optimization and long term forecasts (one year to 10 years ahead) for investment planning. Moreover, load forecasting has become even more important to take part in wholesale markets because demand is the major determinant of the electricity spot prices, at least in France and Germany.

The aim of this paper is to describe the latest forecasting model used at EDF for mid-term load forecasting with a particular focus and experiments on seasonality.

The analysis of the electrical demand shows daily, weekly and yearly cycles reflecting the influence of economic and human activity, calendar effects and weather conditions. These are the main factors to take into consideration in order to build an efficient load forecasting model.

Models used for mid-term load Forecasting at Electricité de France (EDF) therefore decompose the load into two components:

- A part embedding seasonality and trend that is independent from the climate.
- A part dependent on the climate (mostly temperature, and for some model cloud cover).

To achieve an accurate forecast both the weather sensitive part of the load and the seasonal part have to be well estimated. The French electrical load is very sensitive to temperature because of the electrical heating development since the 70's. The influence of the temperature on the French load is mostly known, except for the impact of air conditioning whose trend remains difficult to estimate. Our experience suggests that a better fitting of the seasonality would improve the whole accuracy of the mid-term French load forecasts as well as the short-term forecast when load level changes due to holidays.

We present in section 2 a short literature survey of the methods used to handle seasonality in load forecasting models. As the recent literature is broadly devoted to short term forecasting (STF) we present both the seasonality treatment in STF and mid-term forecasting (MTF). Section 3 presents the MTF model used at EDF with a particular focus on seasonality. In section 4 we present parametric techniques for seasonality modeling, such as regression on different basis functions: trigonometric functions, radial functions or cubic splines. We also present a non parametric model using local regression (LOESS). To deal with the modification of the daily load shape throughout the year, an alternative method was experimented, based on a decomposition into two Fourier series: one depending on the hour and one depending on the day-type. Section 5 addresses the issue of comparing these models. The results are presented in section 6, and conclusions are drawn in section 7.

## 2 LITERATURE SURVEY OF SEASONALITY IN LOAD FORECASTING

Conducting a survey about seasonality in load forecasting models first drives us to a vast literature in the field of economics. Indeed, in order to compare the current observation with that in the previous month, without seasonal effect, economists must apply seasonal adjustment to data. The isolation or the extraction of the seasonal component of economic time series is a difficult issue and is still an active research field [1, 2]. The most common approaches to seasonal adjustment in economics are procedures such as X11 [3] and its successor X12 [4] developed at the US Census Bureau, or the TRAMO/SEATS procedure[5] used for example in the EU statistical office. The X11 and X12 are based on moving averages and TRAMO/SEATS on ARIMA modeling. These methods are designed for quarterly or monthly macro-economic time series like GDP or unemployment. Other methods are also used: regression on seasonal dummy variables, seasonal differences in a Box-Jenkins way and local regression smoothing [6].

Trend and seasonality of the electrical load are quite similar to those of other economic activity series, with an additional effect created by the variation of daylight, throughout the year. Another characteristic of the electrical load is that it exhibits several levels of seasonality: daily, weekly and yearly. This seasonal part is mostly smooth with the exception of daylight savings, which introduce singularities two times a year. The observed seasonality is also affected by economic and human activity, particularly during holidays, and by socio-economic decisions, like "RTT"[1] in France, whose impact is observed but difficult to characterise. Load forecasting models have to take these effects into account. The modeling of the seasonality depends on the forecast horizon and therefore on the kind of load forecasting model. Load forecasting models can be broadly divided in four categories:

- time series models,
- artificial neural networks (ANN),
- similar day look-up,
- regression-based approaches.

The three first ones are mainly STF oriented and the last one is often used in MTF.

### 2.1 Daily and weekly seasonality

To take into account the daily and weekly seasonality, a common approach is to decompose data by day-types, each of them having its own load pattern. The existence of several different day-types has been shown by several researchers. Day-types are determined by both forecasters expertise and clustering techniques. For ANN models[7], inputs include day-types and historical loads, since loads at the same hour are strongly correlated.

For the time series approach, the seasonality is handled by a seasonal ARIMA (SARIMA) [8], or by Kalman filtering[9]. Seasonal differencing using the period of the seasonal variation, usually, 24 (a day) and 168 (a week) for an hourly data set, is required. To introduce an explanatory variable like temperature, SARIMAX models[10, 11] can be used.

In similar day look-up, the similarities lead to a "natural" classification of normal days in five types[12]: {Monday}, {Tuesday, Wednesday, Thursday}, {Friday}, {Saturday} and {Sunday}. For regression based approaches separate models are usually estimated for each day-type. Nevertheless, separate models, per day-type or per-hour, are not always required, and better results can often be obtained by global models having some parameters dependent on the day-types and some other dependent on the hour. The model described in section 3 is of this kind.

### 2.2 Yearly seasonality

In STF models, the estimation of the parameters on a relative short time window allows in most cases to remove the need for yearly seasonality modeling. Another way to address this issue consists in building different models for each season[13]. For example, [14] proposed a seasonal ANN, consisting of 12 independent networks assigned to a particular season and transitions between seasons. In the similar day look-up approach, no particular treatment is required, similar days are likely to be the days in the same season as the day to forecast.

In MTF models, the yearly seasonality can be similarly handled by separate load forecasting models for the different seasons. [15] built a model that only focuses on Summer. [16] used an unsupervised segmentation to break the time series in "winter" and "summer" and learn two separate models for each season. But disaggregation of load by day-types and season can reduce the amount of data available to estimate such models. Therefore, a single model can also be used with dummy variables and trigonometric functions for modeling seasonal effects. [17] used a time-series decomposition and a Fourier series to take into account the yearly seasonality. This kind of model for seasonality is used at EDF and is described in detail in section 3.

## 3 EDF LOAD FORECASTING MODEL

### 3.1 History

EDF has used mid-term regression models since the introduction at the end of the eighties of the nonlinear regression model Météhore[18]. This model, mainly used for French load, has been regularly improved most notably in the PREMIS framework[19] designed by F. Dazy (EDF R&D) in 1997. This model is still in use at EDF. In 2001, a new model, Eventail, based on the same principles was introduced by A. Bruhns and J.S. Roy. This model, used for European countries down to small groups of clients load forecasting and analysis, in various EDF entities, including the EDF Trading subsidiary, is continuously being improved. The current version is described below.

### 3.2 Model Design

Models used for mid-term load Forecasting[2] at Electricité de France (EDF) are regression models based on past values of load, temperature, date and calendar events. The relationship of load to these variables is estimated by nonlinear regression, using a specifically preconditioned variant of S.G. Nash's truncated Newton method[20] developed by J.S. Roy. Load forecasting is performed by applying the estimated model to forecasted or simulated temperature values, date and calendar state. The short term forecasts are performed using an auto-regressive processes applied to the past two weeks residuals of the model.

[1] Reduction of the weekly work schedule from 39 hours to 35 hours.

[2] Also called climatic correction models, since they evaluate the influence of temperature on load.

The model design was strongly oriented toward obtaining a single statistical model (previous models included multiple steps, separating a yearly parameters estimations, and a trend estimation by linear regression). This new model is able to estimate parameters and produce forecasts for all observations, including holidays periods (previous models where unable to cope with holiday periods and invalidated them, requiring the use of specific tools or heuristics to perform a complete forecast). A simple, non heuristic method is used to estimate the parameters (previous models used a mix of linear regression, direct search, smoothing heuristics and manual adjustments). These design decisions simplified implementation, improved performance, and made possible many kinds of performance assessment (confidence intervals on parameters, cross-validation), that were previously very difficult.

### *3.3 Model Description*

The model is based on a decomposition of the load $P_i$, where $i$ indices the observations, into two components:

- $Phc_i$ the weather independent part of the load that embeds trend, seasonality and calendar effects,
- $Pc_i$ the weather dependent part of the load.

The model stipulates[3]:

$$P_i = Phc_i + Pc_i + \varepsilon_i \tag{1}$$

Where $\varepsilon$ is the error of the model. Through the errors are obviously non-gaussian, they are usually assumed to be, purely for simplicity reasons. Correcting this assumption might improve the parameter estimation and the validity of their confidence intervals, but is beyond the scope of this paper.

To describe the model we introduce the following notations:

- $h_i$: the hour for observation $i$,
- $k_i$: the day type:

  The day-types are obtained through clustering, using only predictable calendar information, using a computationally intensive k-means method adapted to the nonlinear nature of the model, since the clustering objective used is to minimize the whole model sum of squared residuals. A bayesian criterion is used to select the number of day-types, usually between 7 and 10, These optimized day-types often offer a forecast quality improvement of approximately 100 MW, about 10% of the RMSE. A perfect but non predictable day-type typology would gain another 100 MW.
- $j_i$: the julian day divided by 365.25,
- $y_i$: the date, in fractional years, from an arbitrarily chosen reference,
- $p_i$: the period of the year:

  We will divide the year in different periods to distinguish:

  - daylight savings in March,
  - holidays in August,
  - daylight savings in October,
  - holidays in December.

  A different value of $p_i$ is associated with each period. During holidays, each week is a different period. At the end, about 10 periods are used.

Whenever non ambiguous, we will drop the indexes $i$ to lighten the notations. In the following description, the estimated model parameters are denoted in **bold**, the indexes specifying that there is one parameter per value of the indexes.

The temperature sensitive part $Pc_i$ is fitted by a non linear model in order to reflect a linear load increase when:

- temperature falls under a cold temperature threshold (heating gradient) ;
- temperature increases over a warm temperature threshold (cooling gradient).

The temperature sensitive part $Pc_i$ is therefore additively separated into two parts, one for heating $\underline{Pc_i}$, and a similar one for cooling $\overline{Pc_i}$, with its temperature threshold reversed. We therefore only present the heating part. Multiple weather stations can be used and each of the weather stations temperature series is exponentially smoothed in the model[4]. Let $t_{i,s}$ be the temperature for observation $i$ and weather station $s$ in $S$, the set of weather stations for which temperature data is available. The smoothed temperature $\underline{u}_{i,s}$ is calculated by:

$$\underline{u}_{i,s} = \underline{\boldsymbol{\theta}}\underline{u}_{i-1,s} + (1 - \underline{\boldsymbol{\theta}})\, t_{i,s}$$

This smoothing is mainly designed to model the inertia in temperature variations inside buildings, with $\boldsymbol{\theta} \simeq 0.98$ for hourly series. The smoothed temperature is then averaged with the observed temperature, to model that a part of the load will be directly related to the observed temperature, and a part of it will be related to the temperature felt inside the buildings. The cloud cover $n_{i,s}$ is added to the resulting value, as a modifier to take into account the greenhouse effect[5] :

$$\underline{v}_{i,s} = (1 - \underline{\boldsymbol{\alpha}}_h)\, \underline{u}_{i,s} + \underline{\boldsymbol{\alpha}}_h t_{i,s} + \underline{\boldsymbol{\mu}}_h n_{i,s}$$

This temperature is an input of a threshold function $\underline{\psi}$ of the form:

[3] The current model includes an additional term to take into account the impact of special tariffs on load.
[4] In the current implementation, the smoothing parameter can be defined per hour.
[5] This last term is absent from the cooling part.

$$\underline{\psi}\left(x, t, \sigma\right) = \mathbb{E}\min\left(x - T, 0\right)$$

Where $t$ is the the temperature threshold, $\sigma$ is the dispersion of the threshold among the population, and $T \sim \mathcal{N}\left(t, \sigma\right)$. See figure 1. This model assumes the temperature threshold $T$ at which heating is started follows a specific distribution whose parameters are to be estimated in the model, and that, below the threshold, the effect of heating on load is linear. This is a strong approximation, and very different effects can be observed for very low temperature, like sharp increases related to electrical supplemental heating, or saturation effects when all heating equipments are already running at their maximum output. Non parametric models are especially effective in this case, but they are more difficult to interpret and to use for stress testing, and are therefore not used here.

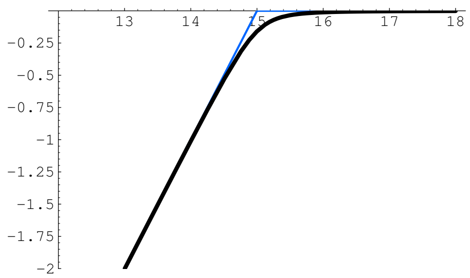


**Figure 1:** Threshold function $\underline{\psi}$ with $t = 15$ and $\sigma = 0.5$.

Finally, the heating part takes the form :

$$\underline{Pc}_i = \sum_{s \in S} \mathbf{g}_{h,s} \cdot \left(1 + \underline{\mathbf{r}} \cdot y_i\right) \cdot \underline{\psi}\left(\underline{v}_{i,s}, \underline{\mathbf{t}}_h, \underline{\boldsymbol{\sigma}}\right)$$

Where $\mathbf{g}_{h,s}$ is the heating gradient for weather station $s$ and hour $h$, $\underline{\mathbf{r}}$ is the gradient's trend, $\underline{\mathbf{t}}_h$ is the heating temperature threshold for hour $h$ and $\underline{\boldsymbol{\sigma}}$ is the heating threshold dispersion. The previously used model averaged all the weather stations temperature before applying a single gradient. This assumes that, for example for a country, all the population reacts to the country's average temperature, while it is quite possible that part of the country are subject to low temperature, while another part is not.

The weather independent component $Phc_i$ is multiplicatively decomposed into three components:

- a load shape part $\mathbf{\Pi}_{h,k}$ that incorporates daily and weekly seasonality,
- a yearly seasonality part $S_i$,
- a trend[6] $R_i = 1 + \mathbf{r} \cdot y_i$.

Thus:

$$Phc_i = \mathbf{\Pi}_{h,k} \cdot S_i \cdot R_i$$

The yearly seasonality $S_i$ is described for each hour by:

$$S_i = \mathbf{q}_{h,p} + F_i \quad (2)$$

The variables $\mathbf{q}_{h,p}$ are dummy variables, per hour $h$ and per period $p$ designed to cope with singularities introduced by daylight savings and holidays. For each hour $h$, a certain amount of load is added or removed depending on the period. The dates of these singularities are known in advance for both daylight savings and holidays.

The part $F_i$ is equal to the first four terms of a Fourier series:

$$F_i = \sum_{m=1}^{4} \mathbf{a}_{h,m} \cos\left(2\pi m j\right) + \mathbf{b}_{h,m} \sin\left(2\pi m j\right)$$

Where $\mathbf{a}_{h,m}$and $\mathbf{b}_{h,m}$ are the coefficients of the Fourier series of the hour $h$. From our experience, using such a model for seasonality makes it unnecessary to use a daylight duration variable in order to take into account the effect of lighting.

Confidence intervals for the parameters are computed using a moving block bootstrap method[21] with a heuristically chosen block size, which may not be a crucial choice, since nearly similar results are obtained for all block sizes greater than a month. These confidence intervals help remove non-identifiable parameters. The implementation of the model and its graphical user interface offer some flexibility, enabling the user to remove irrelevant parts of the model, or to remove degrees of freedom from the parameters, for example by specifying that some parameter (e.g. the heating gradient) must not depend on the hour but only on the weather station. K-fold block cross-validation (see section 5) enables the user to assess the forecasting power (or lack of it) of the estimated model on its data-set.

The current implementation runs on both Unix workstations and Windows PCs. On Windows, a small graphic interface is used to estimate parameters, and an Excel XLL extension is used to perform forecasts and simulations. On Unix, the software makes use of multi-processor workstations for bootstrap and cross-validation tasks. For more intensive testing, a slightly different implementation enable these two tasks to be performed on a cluster of workstations.

While computationally difficult to estimate (strong nonlinearities, particularly with the temperature smoothing parameters and the thresholds, large number of parameters estimated simultaneously, complex estimation function, large data-sets), this model provides reasonably good results, with year ahead forecast errors, for known weather, being around 2%, and day ahead forecasts errors being around 1.5%, without any user intervention.

We nevertheless think that this can be improved. In section 4 we present other basis functions such as radial functions and cubic splines. We also experiment the use of two Fourier series to introduce a relationship with daytypes, and a nonparametric method based on local regression (LOESS). The aim is to provide a model that provides a better fit, especially during holidays.

[6] The trend model is slightly more complicated in practice, allowing for changes in trends.

## 4 METHODS FOR SEASONALITY

In this section we present the models that we have tested to fit the yearly seasonality. In order to deal with daylight savings and holidays, a common approach is the use of dummy variables as presented in 3.3.

Three kinds of models have been experimented:

- Basis function methods: Fourier series, radial basis functions, periodic splines ;
- A non parametric model based on local regression (LOESS) ;
- Sum and product of two Fourier series.

### *4.1 Basis function methods*

The basic model of smoothing procedure in the basis function framework is to fit the data by a function $f$ which is a linear combination of $M$ known basis functions $\phi_m$:

$$f_h(x) = \sum_{m=1}^{M} \beta_{h,m} \phi_m(x) \qquad (3)$$

Different basis functions can be chosen. The best basis should achieve a good fitting of the data with a relative small value of $M$.

We have experimented the Fourier series currently used in the model. Assuming $x$ lies in the $[0, 1]$ interval, the basis functions are: $\cos(2\pi m x)$ and $\sin(2\pi m x)$ where $m$ is an integer value to catch the different frequencies of the data.

We have also tested a periodic cubic splines basis. In a cubic spline regression model, $f$ is constructed by joining together polynomials of degree at most 3. The polynomials are joined at values called knots in such a way that continuity in second derivatives is preserved. The key feature is the number $M$ of knots (equivalent to the number of basis functions) and their location. In our application we used equi-spaced knots. An optimization of the location of the knots would improve the results, but is equivalent to increase the number of parameters. The basis functions are defined by the $M$ splines $\phi_k$, $k = 1, \ldots, M$, where $\phi_k$ is the spline function equals to 1 at knot $k$ and 0 at the other knots.

Finally we have tested radial basis functions. The most used radial basis functions are:

- multiquadric: $\phi(x) = \sqrt{x^2 + c^2}$ for some constant $c$ ;
- inverse multiquadric: $\phi(x) = \frac{1}{\sqrt{x^2+c^2}}$ for some constant $c$ ;
- gaussian: $\phi(x) = e^{-cx^2}$ for some constant $c$ ;
- thin plate spline: $\phi(x) = x^2 \ln x$.

Once the radial basis function $\phi$ has been chosen, the $M$ basis function $\phi_k$, $k = 1, \ldots, M$, are then built by choosing $M$ knots $x_k$, in the definition interval, and defining $\phi_k(x) = \phi(|x - x_k|)$. In our application we used equi-spaced knots. The constant $c$, used in most radial basis functions, is usually estimated separately.

To deal with the daylight savings and holidays the basic model (3) is modified in order to incorporate dummy variables as presented in (2). With the notation of our problem the regression model for the seasonality $S_i$ is:

$$S_i = \sum_{m=1}^{M} \beta_{h,m} \phi_m(j) + \mathbf{q}_{h,p}$$

The next subsection presents a local regression model, which can also be considered as a basis function method, but a nonparametric one.

### *4.2 LOESS*

Local regression (LOESS[22]) is an approach to fit curves and surfaces to data by smoothing. The principle of LOESS is that the fit $f_h(x)$ at $x$ is the value of a parametric function fitted only to observations in a neighbourhood of $x$. The radius of neighbourhood is chosen so that the neighbourhood contains a specified percentage of the data points. The fraction of the data, called the smoothing parameter $\lambda$, controls the smoothness of the estimated surface. Data points in a given local neighbourhood are weighted by a smooth decreasing function of their distance from the center of the neighbourhood, usually the tricube weight function. Weighted least squares are used to fit linear or quadratic functions in the neighbourhood of $x$. In this work we consider a linear function. The smoothing parameter is chosen by cross-validation. In our setting, a LOESS regression is performed per hour $h$, and the values of the dummy variables are estimated by backfitting[23].

$$S_i = f_h(j) + \mathbf{q}_{h,p}$$

### *4.3 Two Fourier Series*

To improve the model and take into account the modification of the daily load shape throughout the year, we have modified the component $F_i$ of the seasonality (2) by introducing another Fourier series. The seasonality becomes:

$$S_i = G_i \cdot (\mathbf{q}_{h,p} + F_i)$$

where $G_i$ is given by:

$$G_i = 1 + \sum_{n=1}^{N} \mathbf{c}_n \cos(2\pi n j) + \mathbf{d}_n \sin(2\pi n j)$$

where $\mathbf{c}_n$ and $\mathbf{d}_n$ are the coefficients of the Fourier series associated with frequency $n$.

We have also experimented a product of a Fourier series per day-type times a Fourier series per hour, by replacing $\mathbf{c}_n$ (resp. $\mathbf{d}_n$) by $\mathbf{c}_{n,k}$ (resp. $\mathbf{d}_{n,k}$), where $\mathbf{c}_{n,k}$ and $\mathbf{d}_{n,k}$ are the coefficients of the Fourier series depending on the day-type $k$.

The last test was to embed in our models a sum of a Fourier series per hour and a Fourier series per day-type. All results of these methods are presented in section 6.

## 5 MODELS COMPARISON

### 5.1 *Extraction of the "noisy" yearly seasonality*

To test the different models of seasonality, three methods were possible:

1. To estimate directly in the current nonlinear model.
2. To approximate the current model with a linear one.[7]
3. To extract the seasonality from the data and to model it apart from the whole model.

The first method is time consuming because of the complexity of the nonlinear model and its estimation. Furthermore, the LOESS method, being nonparametric is difficult to estimate in the current model implementation. We will nevertheless use the current model to test the product of Fourier series because this method requires a nonlinear estimation.

The second and third methods were tested and provided the same conclusion. We only present the third one which is the simplest. The principle is to use the result of the estimated model in order to extract the seasonality component. Combining equations (1) to (2) we obtain:

$$\varepsilon_i = \left( \left( \frac{P_i - Pc_i}{\mathbf{\Pi}_{h,k} \cdot R_i} \right) - S_i \right) \cdot \mathbf{\Pi}_{h,d} \cdot R_i$$

From this equation we can identify $S_i$ which is the estimated yearly seasonality, and $\tilde{S}_i$ which can be regarded as the "noisy" seasonality (see figure 2):

$$\tilde{S}_i = \frac{P_i - Pc_i}{\mathbf{\Pi}_{h,k} \cdot R_i}$$

Our aim is to fit the $\tilde{S}_i$ weighted by $w_i = (\mathbf{\Pi}_{h,k} \cdot R_i)^2$, in order to improve the whole model.

For the numerical experiment, we use an hourly data set for France, with seven years worth of observations from 1995 to 2001. We perform an estimation of the parameters and of all the model's components, and above all the "noisy" seasonality. Performing cross-validation (see subsection 5.2) on this model will obviously underestimate the forecasting error of the whole model, for which many other parameters have to be estimated. Nevertheless, our experiments show that the same number of parameters are obtained by both performing cross-validation on the seasonality only, and on the whole model. This suggests the errors on the estimation of other parameters cannot be corrected by over-fitting the seasonality.

[7] Joint work with J. Collet, EDF R&D.

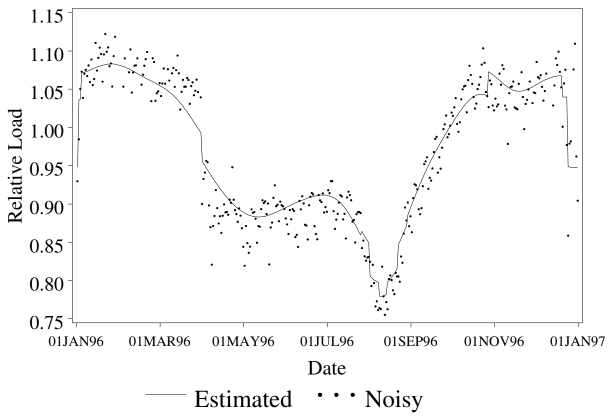


**Figure 2:** Seasonality at 19h, in 1996 : noisy $\tilde{S}$ and estimated $S$.

### 5.2 *Comparison criteria*

In order to compare the models and select the best one, a criterion has to be defined. The main objective of mid term models at EDF is to provide, given a temperature scenario, the most accurate forecast in the RMSE sense, for a one year horizon. It is therefore reasonable to compare the models using the value of the RMSE obtained on such forecasts. The lack of a long enough coherent data set renders difficult its splitting into an estimation and a validation data set. As an alternative, we will use block cross-validation.

Since the data are highly correlated in time, instead of using a classic leave one out cross-validation or k-fold cross-validation, we will chose to leave out a continuous block of observations. The size of the block is empirically guided both by the time dependency of the data and the forecast horizon, usually one year. In our case, we estimate that, beside the yearly periodicity, no correlation should be observable beyond one month. We proceeded to test block sizes of one month and of one year, and got the same results. The following comparisons are therefore done with subsets corresponding to each year. Variance in cross-validation RMSE among the subsets were quite small, so that even small improvements in RMSE may be significant.

In our opinion, the largest bias of this method comes from the fact that we will make forecasts knowing not only the past, but also the future (beyond the period forecast), which is sure to improve the RMSE and bias the model selection towards less robust models. While we are unable to estimate it, we expect that, given the fact that the trends were removed, the bias should be quite small.

Another problem is that the models will probably have many unidentifiable parameters, even if these do not perturb the forecast. While this may interfere with the analysis of the parameters, such analysis is never performed. The fact they have no impact on the forecast is the important point here. Moreover, eliminating these unidentifiable parameters in the model (i.e. nodes for RBF and Spline approaches, and frequencies for the Fourier approach), is quite uneasy, since these parameters might not be the same for each data set.

Nevertheless, this cross-validation approach solves the problem of comparing both parametric and nonparametric models. It would also be possible to compare those models by the number of parameters, by computing an equivalent number of parameters for non parametric models based on the trace of the smoothing matrix. This computation might be affected by the time dependency of the data, and was not used.

## 6 RESULTS

### 6.1 Comparisons inside the current model

Inside Eventail we are able to compare three Fourier approaches: a Fourier series per hour, a single Fourier series per day times a Fourier series per hour, and a Fourier series per day type times a Fourier series per hour.

Our results show that for the Fourier series per hour, the best cross-validation results: 966 MW are attained for 20 terms (i.e., 40 parameters per hour, which seems quite a lot), a much higher value than the 4 terms currently in use, for which the cross-validation error is 991 MW. The seasonality appears much more stable across time than we had expected.

For the product of a Fourier series per day and a Fourier series per hour, the best result is not improved over the use of a single Fourier series. On the other hand, this form appears more parsimonious for a small number of parameters, since with 8 terms for the first series, and 8 terms per hour for the second, this form attains 971 MW, a result only attained with 14 terms per hour for the single Fourier series, i.e., about 6 parameters more per hour.

For the product of a Fourier series per day type and a Fourier series per hour, the best result, 917 MW, was attained for 6 terms in the first series and 20 terms in the second one. A slightly worse result of 925 MW was obtained with 8 terms in the second series. These results suggest that this is clearly the best approach so far. Nevertheless, great care should be taken to ensure that the day types used are similarly distributed among the days of the year for each year of the data set, since the Fourier series per day-type is almost guaranteed to have abnormal values during periods where the corresponding day type were not observed during estimation. This may mandate to use different typologies for the Fourier series and the load patterns used in the model.

We also tried additive models, which gave completely similar results, suggesting that a multiplicative model is unnecessary.

### 6.2 Comparison apart from the current model

In this subsection we present the results obtained for the basis regression methods and the LOESS method, applied to the noisy seasonality described in subsection 5.1, using the same cross-validation criterion. Note that the RMSE values obtained are not directly comparable to the values obtained on the whole model.

For the regression on basis functions, the results of the different methods are comparable : the best RMSE is 938 MW for 32 cubic splines and 32 radial basis (whatever the radial basis function used), and 939 MW for a Fourier series of 16 parameters (equivalent to 32 functions). The results are not so surprising, because in our tests for cubic splines and radial basis functions, the knots were equispaced. The results for these methods would be improved with an optimization of knots location. But this optimization was complex in our cross-validation scheme.

For the LOESS the best RMSE is 948 for a smoothing parameter $\lambda = 0.022$ (a few weeks worth of data around value to estimate). This result suggests that this method is less robust than parametric regression.

## 7 CONCLUSION

This paper describes the latest mid-term load forecasting model used at EDF for mid-term load forecasting. This statistical nonlinear model is able to estimate parameters and produce forecasts for all observations, including holidays periods. Its design makes possible many kinds of performance assessment (confidence intervals on parameters, cross-validation) and provides improved performance. Besides the description of this model, we present some approaches to improve the modeling of seasonality. Parametric techniques such as regression on different basis functions: trigonometric functions, radial functions or cubic splines were experimented. We also present a non parametric model using local regression (LOESS). The results obtained by these methods are comparable to those of the current model. Results would be improved in the case of splines and radial function with knots location optimization.

An alternative model combining two Fourier series, one depending on the hour like in the current model, and one depending on the day-type in order to deal with the modification of the daily load shape throughout the year, was also tested. This last model proves to be the best approach, both in accuracy and parsimony, but requires great care in the day-type typology. This model is now integrated with the current model and performs well: for summer holidays a gain of 25% in RMSE is observed. Further work should address the issue of building different typologies for the Fourier series and the load patterns used in the model. It would also be interesting to experiment alternative approaches to dummy variables.

## References

[1] B. Brendstrup, S. Hylleberg, M.O. Nielsen, L. Skipper, and L. Stentoft. Seasonality in economic models. *Macroeconomic Dynamics*, 8:362–394, 2004.

[2] P.H. Franses and R. Paap. *Periodic Time Series Models*. Advanced Texts in Econometrics. Oxford University Press, 2004.

[3] J. Shiskin, A. Young, and J.C. Musgrave. The X-11 variant of the census method II seasonal adjustment. *Technical Paper 15*, 1967.

[4] D.F. Findley, B.C. Monsell, W.R. Bell, M.C. Otto, and B. Chen. New capabilities and methods of the X-12 ARIMA seasonal adjustment program. *Journal of Business and Economic Statistics*, 16(2), 1998.

[5] V. Gomez and A. Maravall. Automatic modelling methods for univariate series. In D. Pena and G.C. Tiao, editors, *A Course in Advanced Time Series Analysis*, chapter 7. New York: J. Wiley and Sons., fourth edition, 2000.

[6] R.B. Cleveland, W.M. Cleveland, J.E. McRae, and I. Terpenning. STL a seasonal-trend decomposition procedure based on LOESS. *Journal of Official Statistics*, 6(1):3–73, 1990.

[7] H.S. Hippert, C.E. Pedreira, and R.C. Souza. Neural networks for short-term load forecasting : A review and evaluation. *IEEE Transactions on Power Systems*, 16(1):44–55, 2001.

[8] G. Box, G. Jenkins, and G. Reinsel. *Time Series Analysis, Forecasting and Control*. Prentice-Hall, third edition, 1994.

[9] M.-M. Martin. Filtrage de Kalman d'une série temporelle saisonnière, application à la prévision de consommation d'électricité. *Rev. Statistique Appliquée*, XLVII(4):69–86, 1999.

[10] P. Murto. Neural networks models for short-term load forecasting. Master's thesis, Department of Engineering Physics and Mathematics, Helsinki University of Technology, 1998.

[11] F. Meslier, M. Ernoult, R. Mattatia, and P. Rabut. Estimation of the sensitivity of electrical demand to variations in meteorological conditions. Past methods and the development of new approaches by electricité de france. In *Bulletin de la Direction des Etudes et Recherches*, number 3 in B Réseaux Eléctriques, Matériels électriques, pages 5–14. E.D.F., 1981.

[12] J.M. Poggi. Prévision non paramétrique de la consommation electrique. *Rev. Statistique Appliquée*, XLII(4):83–98, 1994.

[13] N.F. Hubele and C.S. Cheng. Identification of seasonal short-term forecasting models using statistical decision functions. *IEEE Transactions on Power Systems*, 1(5):40–45, 1990.

[14] O.A. Alsayegh. Short-term load forecasting using seasonal artificial neural networks. *International Journal of Power and Energy Systems*, 23(3):137–142, 2003.

[15] E.A. Feinberg and D. Genethliou. Load forecasting. In J.H. Chow, F.F. Wu, and J.J. Momoh, editors, *Applied Mathematics for Restructured Electric Power Systems: Optimization, Control, and Computational Intelligence*, pages 269–285. Springer, to appear.

[16] B.J. Chen, M.W. Chang, and C.J. Lin. Load forecasting using support vector machines : A study on EUNITE competition 2001. Technical report, Department of Computer Science and Information Engineering, National Taiwan University, 2002.

[17] E. Pelikan. Middle-term electric load forecasting by time series decomposition. Technical report, Department of Computer Science, Prague University, 2002.

[18] J.P. Ménage, P. Panciatici, and F. Boury. Nouvelle modélisation de l'influence des conditions climatiques sur la consommation d'énergie électrique. Technical Report HR-25/2164, EDF R&D, 1988.

[19] B. Talbot and V. Lefieux. Premis: A new approach to estimate consumption forecasts, from short term to long term. In *Session 38-208*. CIGRE, 2002.

[20] S.G. Nash. Newton-type minimization via the Lanczos method. *SIAM J. Numer. Anal.*, 21:770–778, 1984.

[21] H. Künsch. The jackknife and the bootstrap for general stationary observations. *Annals of Statistics*, 17:1217–1241, 1989.

[22] W.S. Cleveland and C. Loader. Smoothing by local regression : Principles and methods. In W. Hardle and M.G. Schimek, editors, *IEEE Proceedings C*, pages 10–49, 80–102, 113–120. Physica-Verlag, Heidelberg, 1995.

[23] T.J. Hastie and R.J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.