Paper accepted for presentation at 2009 IEEE Bucharest Power Tech Conference, June 28th - July 2nd, Bucharest, Romania

New robust method applied to short-term load forecasting

Yacine Chakhchoukh, Patrick Panciatici
RTE, DMA- Versailles, France
Email: yacine.chakhchoukh@lss.supelec.fr
patrick.panciatici@RTE-FRANCE.COM

Lamine Mili
Virginia Tech-NVC, ECE Department
Falls Church, VA 22043, USA
Email: lmili@vt.edu

***Abstract*—In this paper, the stochastic characteristics of the electric consumption in France are analyzed. It is shown that the load time series exhibit lasting abrupt changes in the stochastic pattern, termed breaks, which need to be accounted for during the modeling process. Thus, a new robust diagnostic approach for which the identification of the breaks is carried out via a robust autocorrelation function estimates is introduced. The developed procedure consists of the following steps: (i) estimate the parameters of a high order autoregressive AR(p^*) by means of the ratio of medians, (ii) execute a robust filter cleaner to reject the outliers, and (iii) apply a maximum-likelihood estimator defined at the Gaussian distribution to handle missing values. The performance of this method has been evaluated on the French electric load time series in terms of execution time, ability to detect and suppress outliers, and forecasting accuracy. The new approach improves the load forecasting quality for "normal days" and presents several interesting properties such as good robustness, fast execution, simplicity and easy on-line implementation. Finally, a simple vector time series method is proposed in order to deal with heteroscedasticity.**

***Index Terms*—Load forecasting, robustness, time series, ARIMA models, median.**

I. INTRODUCTION

The electric load forecasting is a major endeavor carried out on a daily basis by RTE, a company that manages and operates the electric power transmission system in France. It helps RTE to make important decisions regarding the security and reliability of the electric transmission power network. RTE has developed for more than fifteen years a short-term load forecasting method that makes use of autoregressive integrated moving average (ARIMA) models. The method consists of the following steps. The load time series is first corrected from the influence of the weather by using a regression model where the exploratory variables are the temperature and the nebulosity recorded in few selected cities and towns in France. The resulting adjusted series exhibit a general growth trend and three temporal cycles, namely (i) an annual cycle characterized by an annual peak demand in January and a dip on August 15; (ii) a weekly cycle consisting of 5 working days with an overall stable consumption and the weekend when consumption decreases; (iii) a daily cycle. The method of estimation and forecasting implemented at RTE accounts for these cyclic characteristics of the electric consumption.

Next to the meteorological factors and the cyclic features, several parameters influence the electric consumption, which include economical activities, and financial incentives for peak shaving via maximum load reduction tariffs, daylight saving time, and exceptional events. Because of all theses facts, the electric consumption time series encompass a certain number of outliers and few drastic changes in their pattern due to a modification of the mechanisms that generate them, which are caused for example by public holidays. These dramatic changes will be called "breaks" in the sequel.

Fig. 1(a) illustrates the load demand from Saturday July 2^{nd}, 2005 to Saturday July 23^{rd}, 2005. We notice that there is a break appearing on July 14^{th} and lasts until July 17^{th}, 2005 (approximately from observation 600 to 800 on Fig. 1(a)). July 14^{th} is a public holiday in France. These breaks do not follow the general pattern of the time series and hence, affect the classical statistic approach.

A. EXAMPLES OF BREAKS IN THE LOAD TIME SERIES

To fix the ideas, let us give some examples of breaks in the differentiated load time series. Fig. 1(b) displays the one-week differentiated load time series at 7:00 AM in 2005, which is given by $\nabla Y_t^{14} = Y_{t+7}^{14} - Y_t^{14}$ for $t = 1, \ldots, 356$, where Y_t^{14} is the consumption of the day t at 7 AM. We notice the presence of spikes at some sampling times, which are represented by circles in the figure. Called outliers in statistics, these spikes stem from the abrupt changes in the differentiated load series from one day to the other over one week. They are detected by means of diagnostic tools (that is, statistical tests) or are accommodated via robust estimation methods.

Some of the detected breaks are special days such as January 1^{st}, March 28^{th}, May 5^{th}, July 14^{th}, August 15^{th}, October 31^{st}, and November 11^{th} which are non-working days. Note that few special days can not be detected because of their appearance as a weekend. May 1^{st} and 8^{th} are such days.

Obviously, due to the qualitative change observed in the series during the breaks, it is of paramount importance to treat them separately from the majority of the data. It is very difficult and challenging to detect these outliers by experience or eye-balling. This is because of the fact that an observation is judged outlying relative to some model. An observation, which is outlying in some model, may not be outlying in another one. Even if we can pinpoint the public holidays, the starting and the ending time of the public holidays effect on the series cannot be determined with precision.

978-1-4244-2235-7/09/$25.00 ©2009 IEEE

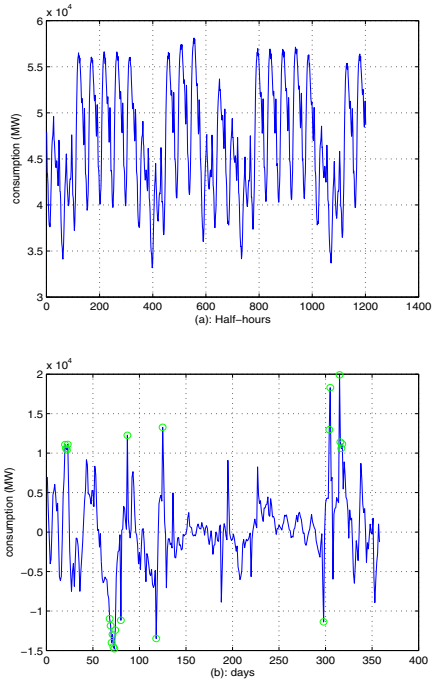


Fig. 1. (a) Half-hourly daily electricity consumption on July 2-23,2005 – (b) One-week differentiated load time series at 7:00 AM in 2005, France

B. ROBUST APPROACH FOR LOAD FORECASTING

To improve the robustness of the parameter estimation and forecasting methods, we may resort to a robust statistical estimation or a diagnostic approach. Good diagnostic approaches achieve robustness via outlier detection and hard rejection, resulting in missing values in the load time series. By contrast, robust methods accommodate outliers by bounding their influence on the estimates, yielding no missing values. The diagnostic approach and robust approach end up to have similar objectives, which are estimating in a robust manner a model. They just do this in a different way.

In the statistical literature, a host of robust estimation methods have been proposed in linear regression, mainly under the assumption of independent and identically distributed (iid) observation errors. On the other hand, for time series applications, there are few methods that have been advocated, which include the filtered τ–, filtered M–, generalized M– and the so-called Residual Autocovariance (RA)-estimators of Bustos and Yohai [1], [2]. Our experience has revealed that robust methods are useful tools for automatic on-line estimation and forecasting load series. These methods can offer a good tradeoff between robustness and efficiency. They constitute also better alternatives to the pragmatic analysis based on experience used by the electric companies. In this article, we propose a novel method termed the ratio-of-medians-based estimator, which is denoted by RME, for short. Some simulation results illustrate the performances of this method in terms of its robustness, forecasting accuracy, and executing time.

The paper is organized as follows. In Section 2, we present the method and analyze its robustness. Section 3 presents the simulation results in the case of electricity load modeling and forecasting. Finally, Section 4 concludes the paper.

II. RME-BASED ESTIMATOR

Our idea is to robustly estimate the autocorrelation and partial-autocorrelation functions, then to fit a high order AR model and filtering with this AR model to "clean" the data from the outliers, and finally to estimate an ARMA model in the presence of missing values.

First, we show how to estimate the correlation ρ in a Gaussian vector using medians, the latter being known to lead to robust estimates. Consider a zero mean Gaussian vector (X, Y) with density

$$\varphi_{\rho,\sigma^2}(x,y) = \frac{\exp\left[-\frac{1}{2\sigma^2(1-\rho^2)}\left(x^2+y^2-2\rho xy\right)\right]}{2\pi\sigma^2\sqrt{1-\rho^2}}. \quad (1)$$

The density of the product XY is given by

$$f_{\rho,\sigma^2}(v) = \int_{-\infty}^{+\infty} \varphi_{\rho,\sigma^2}\left(x, \frac{v}{x}\right)\frac{dx}{|x|} = \frac{e^{\frac{\rho v}{\sigma^2(1-\rho^2)}}}{\pi\sigma^2\sqrt{1-\rho^2}} K_0\left(\frac{|v|}{\sigma^2(1-\rho^2)}\right)$$

where $K_0(\cdot)$ is the modified Bessel function of the second kind [3]. The random variable X^2/σ^2 follows a standard χ^2 distribution of degree 1. For any distribution function F, the median of F is defined as

$$\xi_F = \inf\{x : F(x) \geq 1/2\}. \quad (2)$$

Corresponding to a sample $\{X_1, \ldots, X_n\}$ of observations on F, the sample median $\widehat{\xi}_F$ is defined as the median of the sample distribution function. Let F_{ρ,σ^2} be the cumulative probability distribution function of XY and G_{σ^2} be that of X^2. The ratio τ of medians is expressed as

$$\tau = \xi_{F_{\rho,\sigma^2}}/\xi_{G_{\sigma^2}} \quad (3)$$

$$= \xi_{F_{\rho,1}}/\xi_{G_1}, \quad (4)$$

where $\xi_{G_1} \simeq 0.45$. An explicit relation between τ and ρ does not seem to exist. Fig. 1 depicts ρ as a function of τ, $\rho = r(\tau)$, a relationship that is obtained numerically.

Consider now a Gaussian stationary time series $\{X_t\}$ with variance σ^2 and autocorrelation function $\rho(\cdot)$. For each $k \in \mathbb{N}$, we define $\widehat{\tau}(k)$ as

$$\widehat{\tau}(k) = \widehat{\xi}_{F_{\rho(k),\sigma^2}}/\widehat{\xi}_{G_{\sigma^2}}, \quad (5)$$

where $F_{\rho(k),\sigma^2}$ is the monovariate cumulative probability distribution function of $\{X_t X_{t-k}\}$ and G_{σ^2} is that of $\{X_t^2\}$. The sample medians $\widehat{\xi}_{F_{\rho(k),\sigma^2}}$ and $\widehat{\xi}_{G_{\sigma^2}}$ are calculated using the single ergodic and stationary time series $\{X_t\}$. Then a robust estimate $\widehat{\rho}(k)$ of $\rho(k)$ is obtained by the relation $\widehat{\rho}(k) = r(\widehat{\tau}(k))$. This estimator is called ratio-of-medians-based estimator (RME). To improve the efficiency of the RME, the parameters of an ARMA(p,q) model are estimated

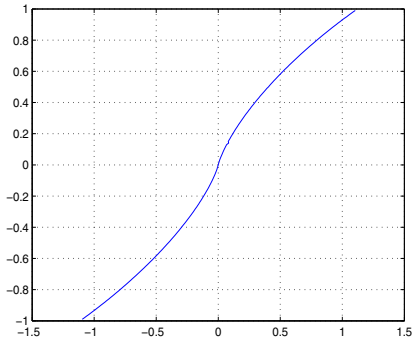


Fig. 2. Correlation coefficient of X and Y versus the ratio of medians, τ, given by (3).

via the following steps:

Step 1: Fit a high order AR(p^*) using the RME, where p^* is selected by a robust order selection criterion subject to being larger than the order of the autoregressive part p.

Step 2: Detect the outliers by filtering with the high order AR(p^*), reject them and use a classical maximum likelihood based estimation method of ARMA models with missing values [4].

The filter replaces the outliers with their expected values obtained from the other observations and the assumed model. While at this stage, we can apply a maximum-likelihood estimator on the 'cleaned' series, we prefer to delete the outliers and apply a classical estimator with missing values [4]. The filter being used is defined in [1] and is based on the robust filter proposed by Masreliez [5], which is termed the filter cleaner. This filter is based on the state representation of an AR(p^*) which is given by

$$\begin{cases} X_t = \Phi X_{t-1} + D\varepsilon_t \\ y_t = GX_t \end{cases}, \tag{6}$$

$$\Phi = \begin{pmatrix} \phi_1 & \\ \vdots & I_{p^*-1} \\ \phi_{p^*} & 0'_{p^*-1} \end{pmatrix} \tag{7}$$

Here, Φ is the transition matrix,$D = (1, 0, \ldots, 0)'$, $G = (1, 0, \ldots, 0)$, I_k is the $k \times k$ identity matrix and 0_k the zero vector in $\mathbb{R}^k$;dim(Φ)=$k \times k$.

A. Maximum bias curves in the case of AR(1)

The maximum bias curves of the RME are calculated following the Monte Carlo procedure described in [1, page 305]. For AR(1), Fig. 3 depicts the maximum bias curve of our RME together with that of another robust estimator, namely the GM estimator. It is observed from these plots that RME is robust and has a breakdown point of about 25%, that is, it can handle up to 25% of outliers among the data samples. The breakdown point of 25 % observed in simulations can be explained as follows. Because in our model, y_p affects two terms of the product $y_{p-1}y_p$ and y_py_{p+1}, the breakdown point of the RME is half that of the sample median, yielding BP=1/4. Note that while the GM estimator seems to perform better for a small fraction of contamination, its robustness degrades with increasing dimensions, signifying a decreasing breakdown point with increasing number of parameters to be estimated, a well-known result in robust statistics literature (i.e. [1]). On the other hand, our RME exhibits a constant breakdown point regardless of the order of the AR model.

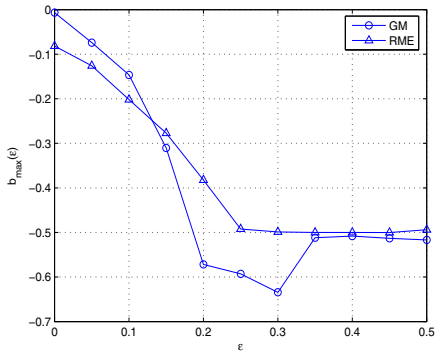


Fig. 3. Maximum bias curves of two robust estimators of an AR(1) , ϕ=0.5

B. Modeling the load time series

The final goal is to robustly fit an ARIMA model. This ARIMA will be used to forecast the "normal" days. The breaks are treated and forecasted separately from this model. A seasonal ARIMA model, SARIMA $(p^h, 0, q^h) \times (p_1^h, 1, q_1^h)_7$ follows the equation

$$\phi^h(B)\Phi^h(B^7)\nabla_7 Y_t^h = \theta^h(B)\Theta^h(B^7)\varepsilon_t^h,$$

where Y_t^h is the daily electricity demand on day t at time h ($h = 1, \ldots, 48$, at 12:00 $h = 24$), the number of periods is equal to 7 to model the within-week seasonal cycle. B is the lag operator. ∇ is the difference operator, ∇_7 is the seasonal difference operator ($B^l Y_t^h = Y_{t-l}^h, \nabla = 1-B, \nabla_7 = 1-B^7$). ϕ^h, Φ^h, θ^h, Θ^h are polynomials of order p^h, p_1^h, q^h, q_1^h. ε_t^h is a Gaussian white noise from $N(0, \sigma_\varepsilon^2)$.

Casted in a multivariate modeling framework, the electric consumption is represented as a set of 48 time series corresponding to the hours of the day with a statistical model for each hour (00:00, 00:30,. . ., 24:00) and considering the correlation between adjacent hours.

The residuals obtained from the ARIMA model at various hours are standardized via a robust estimator of scale. They are referred to as the standardized residuals. The standardized residuals of adjacent hours are correlated and their correlation is modeled by an ARMA model defined as

$$\varphi(B)r_t^h = \vartheta(B)\varepsilon'_{48(t-1)+h} \tag{8}$$

This model is used to improve the prediction. The series of the residuals is given by $r_1^1, \ldots, r_1^{48}, \ldots, r_n^1, \ldots, r_n^{48}$, where

n is the number of days in the series and r_k^h is the residual of the day k for the series of hour h. The idea is to use the correlation between adjacent hours in the prediction without employing a too complex vector time series approach, which is difficult to develop and implement.

III. Simulation results on the French load series

The dataset consists of intraday hourly load from February $1^{st}, 2004$ to June $14^{th}, 2005$. The mean-absolute percentage error, MAPE, which is defined as in Eq. (9), is computed over 100 days (from Mars $6^{th}, 2005$ to June $14^{th}, 2005$). It is a robust index of performance since it is computed over normal days. This makes sense since the principal goal of robust estimation and forecasting is to increase the forecasting quality of the majority of the data. We use an on-line and an off-line estimation and forecasting procedure. The off-line procedure estimates one (ARIMA) model to carry out the forecast. On the other hand, the on-line procedure estimates several models and adapts the ARIMA estimation to newly recorded data. For the first forecast, the first 500 days are used to estimate the parameters while the remaining 100 days are employed to evaluate post-sample accuracy of forecasts up to ten days ahead for all the hours of the day. For the next forecasts, the same procedure is applied. Fig. 4 depicts the locus of the MAPE, which is given by

$$\text{MAPE} = \frac{100}{l} \sum_{t=1}^{l} \left| \frac{Y_t^h - \widehat{Y}_t^h}{Y_t^h} \right|, \tag{9}$$

where l is the length of prediction. $\widehat{Y}_t^h$ is the predicted value at day t. The maximum-likelihood-based classical approach applied after smoothing the unusual observations by three-sigma rejection rule, is denoted by CML. The three-sigma rule consists in treating data that are outlying beyond three times the standard deviation from a robust estimate of the trend (that is, the central part) of the time series. The mean-absolute-percentage error of the CML and of our robust first approach are denoted by MAPE and MAPER, respectively. We notice that for all leading times, MAPER $<$ MAPE. This means that the forecasting quality is improved with the new approach for these two hours of the day.

Fig. 5 depicts the MAPE obtained by the on-line RME-based estimator and the on-line Generalized-M estimator. It is seen that the RME based estimator offers better performance than the GM estimator. This can be explained by the deteriorating performance of the GM estimator when the dimension increases, which is the case for the high order autoregressive model. We also notice that the on-line RME-based estimation has better performance than the off-line RME. We conclude that the RME outperforms the others for on-line applications thanks to its simplicity, robustness and small computing time.

In Fig. 6, we show the quantiles of the absolute values of the residuals of two estimates at the series of hour 10:00. Namely, the proposed RME-based estimate (MAPER) and the classical approach, denoted by CML, in the French daily load forecasting (MAPE). It is seen that the RME-based estimate yields the smallest quantiles, and hence gives the best fit to the bulk of data. We remark that a small fraction of residuals obtained with the RME-based estimator are very large. These residuals correspond to the outliers.

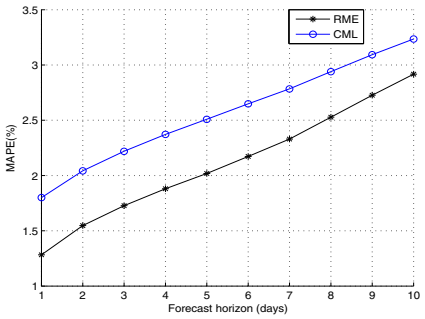
A good criterion to compare the forecasting quality of different models obtained is the error between the measure at time k and the forecast of this value at time k-1 for example (ie. $\widehat{Y}_{k|k-1}$). We compute the one-hundred first-step forecast errors for the two models obtained with the RME method and the CML method. We estimate the median and the MADN $=$ median$(|\widehat{e}_{t|t-1} -$ median$(\widehat{e}_{t|t-1})|)$ and calculate the Robust Mean Squared Error (RMSE $=$ median^2 + MADN^2) for the two previously mentioned estimators. At 12:00, we obtain 5.3598×10^5 and 11.459×10^5 for the RME and the CML, respectively.

In Fig. 7, the MAPE for hours 8:00 and 22:00 is presented. We remark that the improvement of the forecast varies from hour to hour. This is quite natural since some hours are more contaminated than others with outliers. The time series in the night hours is not contaminated by non-working days for example.

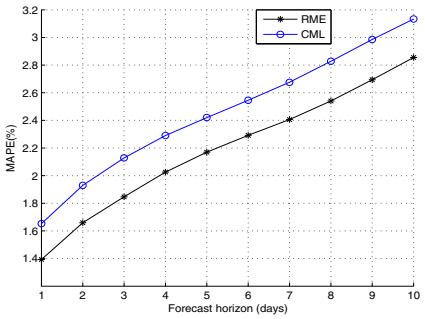
In Fig. 8, we show the mean absolute percentage error of the RME-based estimator and the classical maximum likelihood estimator for all the hours of day by using the approach explained in Section II-B. It is clear from the curves that the RME-estimator improves the quality of forecasting. In the time interval that ranges between the 5^{th} hour (2:30) and 13^{th} hour (6:30), the classical maximum likelihood have better performance than the RME-based estimator. This can be explained by the fact that this hours, being late in the night (2:30-6:30), does not contain outliers. In this case, the best estimator is the maximum likelihood estimator since it is the most efficient estimator at the Gaussian distribution.

IV. Conclusion

In this paper, a new robust method for ARIMA models estimation is proposed, namely the RME method. We compare its performance to that of the classical approach based on the maximum likelihood estimation, which is applied to the French daily load forecast after carrying out a smoothing by the three-sigma rejection rule. Furthermore, a new modeling approach relying on vector time series is proposed. It is found that the robust method outperforms the classical methods. The RME is the best method for this application regarding its simplicity, time of execution and robustness. Ongoing effort has been concentrating on computing the asymptotic distribution and variance of the RME when it is applied to a correlated series. This analysis will allow us to derive confidence intervals for the autocorrelations and possibly for the parameters as well. Furthermore, a comparison of the RME with the median of slopes estimator, which has bias-optimality properties [1, Chapter 5], is worth investigating. Another research work will be to compare the performance of the RME to that of the τ-estimator and the MM-estimator in load forecasting.



(a) 12:00



(b) 18:00

Fig. 4. MAPE forecast accuracy versus lead time for the series at 12:00, 18:00 with off-line estimation

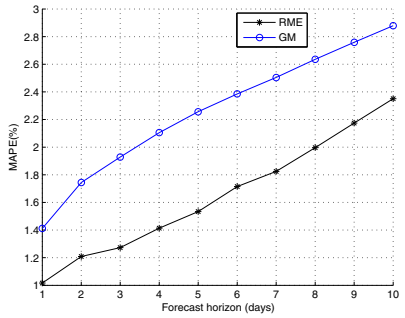


Fig. 5. MAPE forecast accuracy versus lead time for 12:00 in the case of online estimation

REFERENCES

[1] Ricardo A. Maronna, R. Douglas Martin, and Victor J. Yohai, *Robust statistics*, Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2006, Theory and methods.

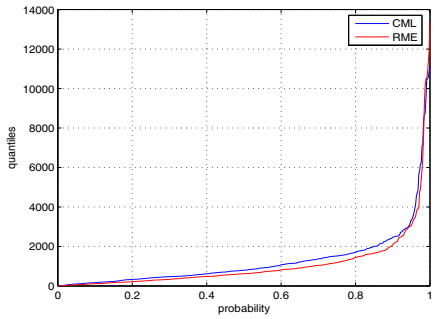
[2] Oscar H. Bustos and Víctor J. Yohai, "Robust estimates for ARMA models," *J. Amer. Statist. Assoc.*, vol. 81, no. 393, pp. 155–168, 1986.

[3] Milton Abramowitz and Irene A. Stegun, Eds., *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, Dover Publications Inc., New York, 1992, Reprint of the 1972 edition.

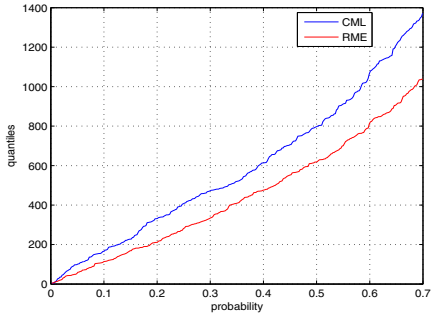
[4] Richard H. Jones, "Maximum likelihood fitting of ARMA models to time series with missing observations," *Technometrics*, vol. 22, no. 3, pp. 389–395, 1980.

[5] C. Johan Masreliez and R. Douglas Martin, "Robust Bayesian estimation for the linear model and robustifying the Kalman filter," *IEEE Trans. Automatic Control*, vol. AC-22, no. 3, pp. 361–371, 1977.

[6] Rafal Weron, *Modeling and Forecasting Electricity Loads and Prices: A Statistical Approach (The Wiley Finance Series)*, Wiley, December 2006.

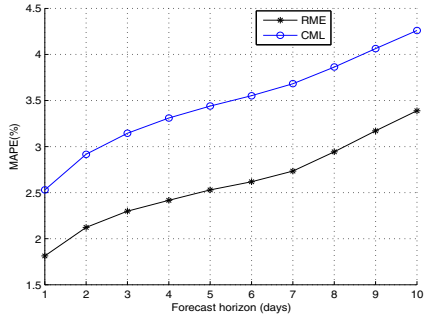


(a) 10:00

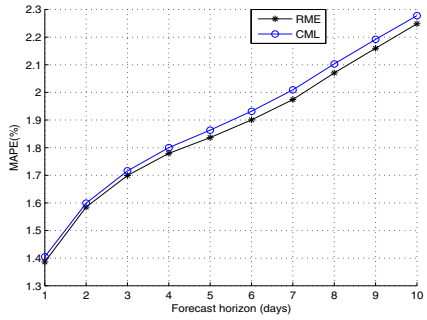


(b) 10:00

Fig. 6. Quantiles of absolute residuals of estimates for load series at 10:00



(a) 08:00



(b) 22:00

Fig. 7. MAPE forecast accuracy versus lead time for the series at 08:00, 22:00 with off-line estimation

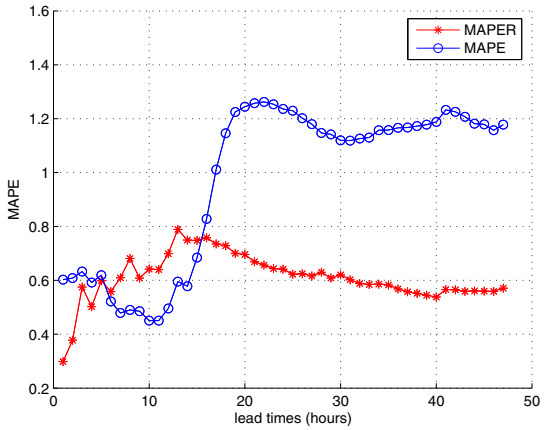


Fig. 8. MAPE forecast accuracy versus lead time for all hours of the day using on-line estimation

[7] James W. Taylor and Patrick E. McSharry, "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.

[8] Shyh-J. Huang and Shih Kuang-R., "Short-term load forecasting via arma model identification including non-gaussian process considerations," *IEEE Transactions on Power Systems*, vol. 18, no. 2, pp. 673 – 679, 2003.

[9] Lacir Soares and Marcelo C. Medeiros, "Short-term load forecasting methods: An evaluation based on european data," *IEEE Transactions on Power Systems*, vol. 22, pp. 2213–2219, 2007.

V. BIOGRAPHIES

Yacine Chakhchoukh was born in Algiers, Algeria, on May 1982. He graduated from Ecole Nationale Polytechnique d'Alger and he is preparing a Ph. D. in Signal Processing and Load Forecasting at Réseau de Transport d'Electricité (RTE, the French transmission system operator) and the Laboratoire des signaux et systèmes, CNRS-Supélec-Université de Paris-Sud XI.

Patrick Panciatici was born in March 1960, graduated from Ecole Supérieure d'électricité Supélec in 1984. He joined EDF Research in September 1985. Since 1998, he is the head of a team which develops Security Analysis tools for real time (EMS) and operational planning now at RTE (French transmission system operator). During the last 17 years, he has contributed in different projects at EDF Research : Coordinated Secondary Voltage Control (CSVC), Power System Simulation (EUROSTAG) and Short Term Load Forecast.

Lamine Mili received an Electrical Engineering Diploma from the Swiss Federal Institute of Technology, Lausanne, in 1976, and the Ph. D. degree from the University of Liege, Belgium, in 1987. He is presently a Professor of Electrical Engineering at Virginia Tech. His research interests include state estimation, transient stability, voltage collapse, and power system control.