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

Confidence interval estimation for short-term load forecasting

B. Petiau, RTE, Réseau de Transport d'Electricité, Power System Department, Methods & Support

Abstract--**This paper presents a method to obtain confidence intervals (CI) for load forecast. It is based on the calculation of empirical quantiles of relative forecast error observed in the past. A classification is made between days for which the load forecast is difficult and those for which it is easier ; this classification is made a priori, based on forecast error knowledge. Some CI's skills for the confidence interval are evaluated on a test period and show this method is more accurate and useful than a basic method simply based on the standard deviation of error. CIs provide a way of quantifying the uncertainty of the forecast. For TSOs, application fields are numerous. They could be used to assess as precisely as possible the operating margins. They could also be used to generate extreme demand scenarios, at a given risk level, for security analysis studies.**

***Index Terms*—Error analysis, Load forecasting, Power demand**

I. Introduction

SHORT-TERM load forecasting (STLF) is a major issue for RTE, the French Transmission System Operator (TSO). On a daily basis, RTE performs day-ahead forecasts for the French demand and for local French areas' demands. These forecasts are used to guarantee power generation versus consumption balance at any moment of the next day. They are also essential for daily network studies (achieved both on national and regional scales) analyzing electricity security, congestions, and so on.

From a statistical viewpoint, the STLF problem is complex for two reasons : 1) the relation between load and exogenous factors - time factors, weather conditions – is non truly linear 2) the load time series is disturbed by "breaks" (winter holiday break, public holidays).

STLF methods can be divided into two categories. Some models are based on machine learning methods : Artificial Neural Networks [1], Fuzzy Logic, or Support vector Machines. On the other hand, statistical methods represent loads as a function of different factors : exponential smoothing regression methods, time series models. An overview of all these methods is presented in [4]. RTE day-ahead forecasting method is a statistical model decomposing demand into two parts. Firstly, a part sensitive to the climate, we could think of usage such as heating and air conditioning, which is modelized by non linear and non stationary transfer function with temperature and nebulosity as input variables. For day-ahead, the model uses temperature and nebulosity forecasts. Secondly, endogenous factors like seasonality effects (days, week) are modelled with a SARIMA model.

Today, the mean performance of RTE load forecasting model (e.g. Mean Absolute Error (MAE) value) is not at stake. The main concern is extreme errors that may occur during cold spells, holidays or vacation breaks. To cope with this issue, forecasters need a system to quantify load forecast uncertainty. In this paper, we propose a statistical system of confidence intervals (CIs) for the STLF problem, which is a first step towards full probability density function forecast. Let $\hat{L}_{d_0,h/(d_0-1)}$ be the day ahead forecast for the load $L_{d_0,h}$ at hour h of day d_0 made at d_0-1; a $(1-\alpha)100\%$ CI is an interval $\left[L^{\inf,\alpha}_{d_0,h/(d_0-1)} \quad L^{\sup,\alpha}_{d_0,h/(d_0-1)} \right]$ such as : $P\left(L_{d_0,h} \in \left[L^{\inf,\alpha}_{d_0,h/(d_0-1)} \; L^{\sup,\alpha}_{d_0,h/(d_0-1)} \right]\right) = 1-\alpha$. Note the construction of a confidence interval is similar to the derivation of the probability law of forecast errors : for example, $L^{\inf,\alpha}_{d_0,h/(d_0-1)}$ is the $(\alpha/2)100\%$ quantile of $L_{d_0,h}$: $P(L_{d_0,h} < L^{\inf,\alpha}_{d_0,h/(d_0-1)}) = \alpha/2$.

Due to the inhomogeneity of forecasting errors, theoretical CI for the RTE load forecasting doesn't give accurate information of extreme errors during holidays and vacation breaks. Consequently, we have to estimate an empiric system of CI. There are only a few references on the estimation of CIs for STLF. There is no general methodology of CIs estimation, because the interval construction depends on the type of model used for the forecast (for instance, parametric, ANN, non parametric, decision tree). [1] and [2] present such CI systems.

In section 2, we present variables that are used for qualifying the possible range of forecast errors. We propose next a specific method of CIs construction for the RTE model. The principle of this method is very simple and could be easily adapted to other STLF models. In section 3, we show the results obtained on real data, and we discuss about means of evaluating confidence interval quality. The model presented in section 2 is compared to simpler a model used as a reference. Section 4 propose some perspectives of applications of CI's for SLTF. We conclude in section 5 with possible extensions of the proposed model.

In the entire article, applications are done for the day-ahead load forecast of a French region.

This work was supported by Réseau de Transport d'Electricité (Power System Department, Methods & Support).

B. Petiau is with Réseau de Transport d'Electricité, Power System Department, Methods & Support (e-mail : Benjamin.Petiau@rte-france.com)

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

II. TECHNIQUE FOR CI ESTIMATION

A. Analysis of day-ahead load forecast error

The estimation of the CI for the day $d_0, d_0 > T$ is based on the statistical observation of relative forecast errors of the past $\varepsilon_{d,h/(d-1)} = \left(L_{d,h} - \hat{L}_{d,h/(d-1)}\right) / \hat{L}_{d,h/(d-1)}$ for a set of days $d \in [0\,T]$ in an estimation period of the past. We therefore assume that the empirical distribution of relative forecast errors is not significantly modified in time i.e. the relative forecast errors that have been in the past can be used for the relative forecast error for the future.

Our aim is to build a confidence interval that is changing each day depending on the situation. Before building a model, it is important to find explanatory variables that are relevant to explain the probability law of relative forecast errors.

Because the quality of forecast varies a lot during the day, the forecast hour is the first variable used for modelling the variability of the error. For example, Fig. 1. shows the empirical distribution of relative forecast error for two different hours in a day. Empirical distribution are computed with a kernel smoothing density estimate. A two-sample Kolmogorov-Smirnov test rejects the hypothesis that error at 1:00 and error at 19:00 have the same continuous probability distribution function.

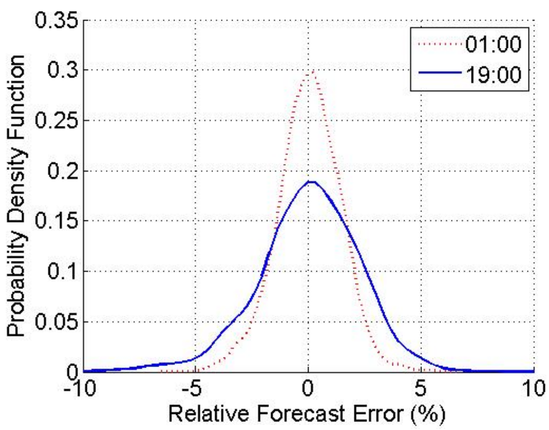

Fig. 1. Distribution of relative forecast error for two different hours in a day

Secondly, some periods in a year are easier to forecast than others. For example, Fig.2. shows the difference between distribution of errors for June and January.

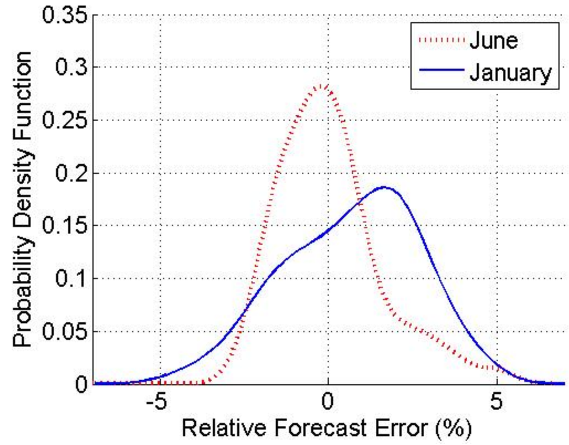

Fig. 2. Distribution of relative forecast error for 2 different months

We have done an *a priori* classification of days based on error forecasts knowledge. Each day is classified into one of the following clusters :

Cluster 1 : normal days
Cluster 2: winter holiday break
Cluster 3: days preceding a public day
Cluster 4: public days
Cluster 5: days following a public day

As for hour and period of year, we compute the error distribution of each cluster, that is represented on Fig.3. Table 1 shows p-value of the different Kolmogorov-Smirnov tests of equal distributions for each cluster. When the p-value is low, the hypothesis that error of the two clusters come from the same distribution is rejected. We conclude from results of table I that errors from each cluster globally come from different distributions. The only exception is cluster 5, that could be associated with cluster 3 : errors of days preceding a public day and for days following a public day could be considered from the same continuous distribution.

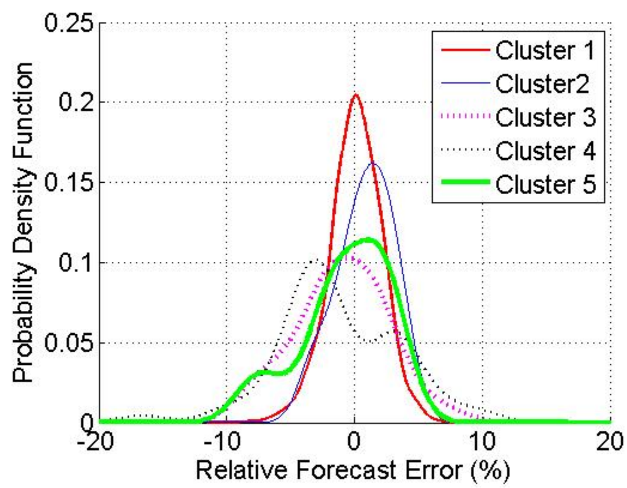

Fig. 3. Distribution of relative forecast errors for the five different clusters

TABLE I P-VALUE OF KOLMOGOROV-SMIRNOV TEST BETWEEN ERROR OF CLUSTERS

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Cluster 1	1.00	4.6%	0.1%	0.0%	6%
Cluster 2		1.00	0.9%	0.0%	12%
Cluster 3			1.00	26%	78%
Cluster 4				1.00	4.5%
Cluster 5					1.00

B. CI's computation method

Confidence interval method is directly computed from observations of subsection A. Let us recall that we can use a base of statistical observations of relative forecast errors of the past on a set of days $d \in [0;T]$ named estimation period. Our aim is to build confidence interval for the load forecast of a day d_0 ($d_0 > T$) for hour h. Let g_0 be the cluster of the day d_0. There are two cases, depending of g_0:

- If $g_0 = 1$ (normal days), we select the set $E\{d_0\}$ of dates in $[0;T]$ that are in the same period of year than d_0 and that belong to cluster 1. The definition of a period is plus or minus 20 days centered on the same calendar date as d_0 in the past.
- If $g_0 \neq 1$ (cluster 2 to 5), $E\{d_0\}$ is the ensemble of dates in $[0;T]$ that belong to the cluster g_0. Because there are not a lot of observations in each of the clusters 2 to 5, we do not use the calendar criteria selection as for normal days.

This procedure produces a set of days $d \in E\{d_0\}$ such that for each hour h errors $\varepsilon_{d,h/(d-1)}, d \in E\{d_0\}$ are supposed to follow the same probability law than that of the error $\varepsilon_{d_0,h/(d_0-1)}$. Then, CIs for the load $L_{d_0,h}$ at hour h of day d_0 are constructed with the empirical Quantiles $q_h^{\frac{\alpha}{2}}(E\{d_0\})$ and $q_h^{1-\frac{\alpha}{2}}(E\{d_0\})$ on relative errors $\varepsilon_{d,h/(d-1)}, d \in E\{d_0\}$. Consequently, the lower and upper limits of the CIs have the following form :

$$\begin{cases} L_{d_0,h|(d_0-1)}^{\inf,\alpha} = \hat{L}_{d_0,h|(d_0-1)}\left(1+q_h^{\frac{\alpha}{2}}(E\{d_0\})\right) \\ L_{d_0,h|(d_0-1)}^{\sup,\alpha} = \hat{L}_{d_0,h|(d_0-1)}\left(1+q_h^{1-\frac{\alpha}{2}}(E\{d_0\})\right) \end{cases} \quad (1)$$

Note that this procedure is done independently for each hour of the day, so the length of the confidence interval depends on the value of the forecast $\hat{L}_{d_0,h/(d_0-1)}$, the cluster of the day d_0, the hour of the day h, and the period of year (for day in cluster 1). Hence, in the following, this model will be named "adaptive CI's method".

III. RESULTS AND TESTS

A. Results

The method has been implemented in a software that is being experimented by RTE forecasters. For example, Fig. 4 shows the CI calculated for a normal day (Cluster 1) in the winter for the load forecast of a French region. The user can specify the level of confidence. Fig. 4 shows an *a posteriori* analysis with the actual load appearing within the CI. Forecasts could be calculated for the whole French network or for each of the seven French regional network. Results can vary a lot from a region to another.

The method has been calibrated on a four years period and tested on a one year and half period. Fig. 5 shows the evolution of load, load forecast, and confidence interval on the test period. The variability of the length of confidence interval is more detectable on Fig.6 that shows the evolution of load forecasting error $\left(L_{d,h} - \hat{L}_{d,h|d-1}\right)$ with the 90% associated confidence interval. Note that the width of confidence interval increase in winter and is larger for public days. That follows the behavior of error statistically described in section II.

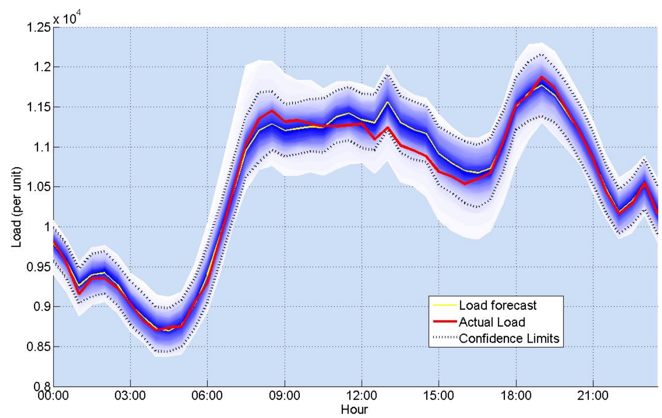

Fig. 4 : Example of 90% confidence intervals for a normal day in the winter

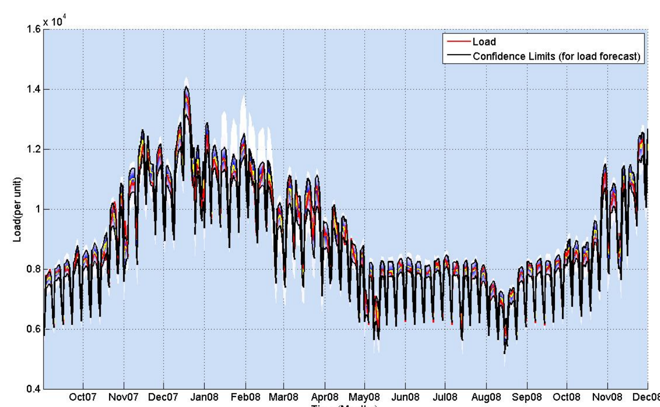

Fig. 5 : Evolution of load with associated confidence interval

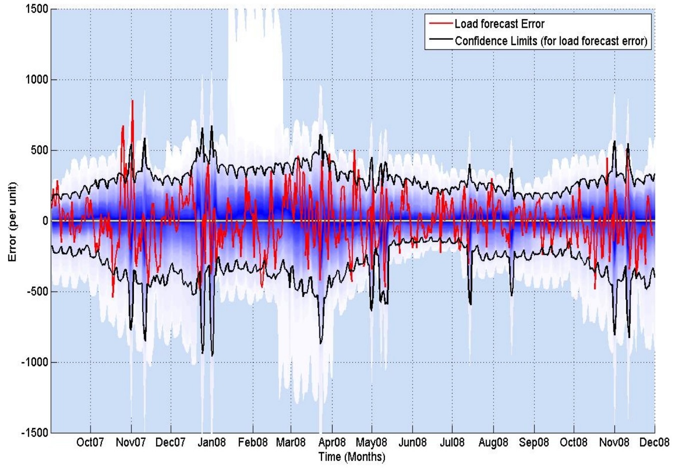

Fig. 6. Evolution of load forecasting error

B. Tests

1) Basic Method

In B.2) we will define and evaluate some CI's skills for the confidence interval on the test period. To compare the performance of the adaptive CI's method described in section II), we need to compute another method, much simpler, that will be used as a reference. Let us define this very simple confidence interval method : for each hour h, we suppose the absolute error $\varepsilon_{d,h/(d-1)} = \left(L_{d,h} - \hat{L}_{d,h/(d-1)}\right)$ follows an i.i.d Gaussian probability law, the basic confidence intervals is composed by multiples of the standard deviation of $\varepsilon_{d,h/(d-1)}$. Fig. 7 shows the evolution of load forecasting error $\left(L_{d,h} - \hat{L}_{d,h|d-1}\right)$ with the 90% associated confidence interval for the basic confidence interval. Clearly, the width of interval is constant and is not adapted for each situation.

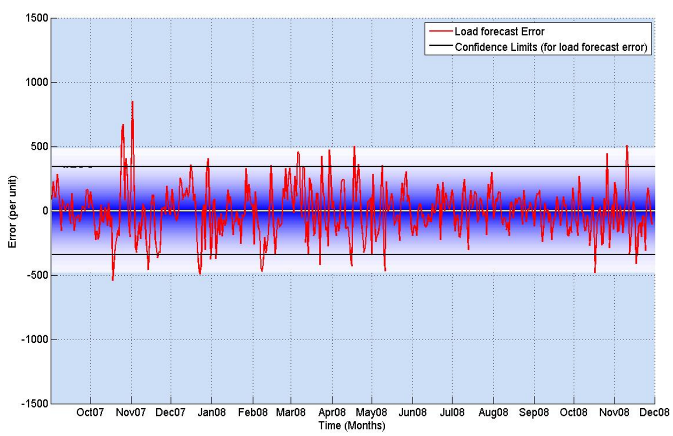

Fig. 7 Evolution of load forecasting error for basic confidence interval method

2) Tests

We have evaluated some CI's skills for the confidence interval on the test period. First of all, the confidence interval must be accurate. We name *Coverage* the proportion of times confidence intervals $\left[L^{\inf,\alpha}_{d_0,h|(d_0-1)} \; L^{\sup,\alpha}_{d_0,h|(d_0-1)} \right]$ containing the "true" load $L_{d_0,h}$ for d_0 in the test period. Let's N be the number of observations in the test period. The empiric coverage $C_h(\alpha)$ associated to the level $1-\alpha$ can be expressed as :

$$C_h(\alpha) = \frac{\sum_{d_0 \in T} I_{d_0}(\alpha)}{N} \qquad (2)$$

$$\text{with } I_{d_0,h}(\alpha) = \begin{cases} 1, \, if \, L_{d_0,h} \in \left[L^{\inf,\alpha}_{d_0,h|(d_0-1)} \; L^{\sup,\alpha}_{d_0,h|(d_0-1)} \right] \\ 0, \, if \, L_{d_0,h} \notin \left[L^{\inf,\alpha}_{d_0,h|(d_0-1)} \; L^{\sup,\alpha}_{d_0,h|(d_0-1)} \right] \end{cases}$$

The closer the coverage is to $1-\alpha$, the more accurate is the confidence interval. References [3] and [8] propose a statistical test for $C_h(\alpha)$ examining the hypothesis $H_0 : E(I_{d_0,h}(\alpha)) = 1-\alpha$ against $H_1 : E(I_{d_0,h}(\alpha)) \neq 1-\alpha$. If H_0 is not rejected, confidence should be considered as valid. Supposing the indicator variable $I_{d_0,h}(\alpha)$ has a Bernoulli distribution, the likelihood function under the null hypothesis is :

$$L(\alpha) = \alpha^{n_{missed}} (1-\alpha)^{n_{hit}} \qquad (3)$$

whereas under (H1) the likelihood function is :

$$L(C_h(\alpha)) = C_h(\alpha)^{n_{missed}} (1 - C_h(\alpha))^{n_{hit}} \qquad (4)$$

where $n_{hit} = \sum I_{d_0,h}(\alpha)$ and $n_{missed} = N - n_{hit}$.

Then, the coverage is tested with the likelihood ratio test ; supposing H_0 is true, there is :

$$LR = -2\ln(L(\alpha)/L(C_h(\alpha))) \sim \chi^2(1) \qquad (5)$$

Then, p-value of the test is :

$$p\ value = 1 - F_{\chi^2(1)}(LR) \qquad (6)$$

where $F_{\chi^2(1)}$ is the $\chi^2(1)$ cumulative distribution function.

Table II shows the mean coverage with in brackets the associated p-value of the likelihood ratio test at different hours. p-value higher than 1% means that the mean coverage is not significantly different from the expected confidence level $1-\alpha$. Results for adaptive method (described in section II) would be compared to those of the basic method. In most cases, empiric coverages are close to the expected confidence level for confidence intervals constructed with adaptive method, and high p-value allow us not to reject the null hypothesis ; we conclude on the validity of confidence intervals calculated with the adaptive method.

TABLE II : MEAN EMPIRICAL COVERAGE OF CONFIDENCE INTERVAL AND ASSOCIATED LILELIHOOD RATIO TEST P-VALUE

	Confidence Level (expected)	80%	90%	95%
Adaptive method	7:00	84% (0.7%)	91%(22.9%)	95%(87%)
	13:00	79% (64%)	92%(2.6%)	95%(87%)
	19:00	79% (48%)	89%(58%)	95%(95%)
Basic method	7:00	90% (0%)	94%(0.1%)	96% (0.2%)
	13:00	85% (0..2%)	92% (8%)	96% (29%)
	19:00	84% (2%)	91% (22%)	95% (79%)

The second criteria to assess the quality of confidence interval is the relative witdth defined as :

$$\delta L_h = mean\left(100\frac{\left(L^{\sup,\alpha}_{d_0,h|(d_0-1)} - L^{\inf,\alpha}_{d_0,h|(d_0-1)}\right)}{2L_{d_0,h}}\right)$$

Tab 3 indicates values of δL_h for different hours in the day. Clearly, its value depends on the hour of day and increases with the expected confidence level. Note that the relative width is quite lower for the adaptive model. For applications like margin's assessment, it is important to provide CIs with the lowest relative width because RTE's operators can not use a too big load hypothesis.

TABLE III : RELATIVE WIDTH OF CONFIDENCE INTERVALS

	Confidence Level (expected)	80%	90%	95%
Adaptive model	7:00	2.8%	3.6%	4.3%
	13:00	2.3%	3.0%	3.6%
	19:00	2.7%	3.6%	4.3%
Basic Model	7:00	3.1%	4.0%	4.7%
	13:00	2.6%	3.4%	4.0%
	19:00	3.1%	4.0%	4.7%

The last CI skill we have evaluated is the capacity of confidence intervals to discriminate situation with a high risk of errors from those with a low risk. To evaluate this capacity, we propose a criteria very close to the area under the ROC (Receiving Operational Curve) curve which is well known in signal detection theory (see for example [5] or [6] for more details on ROC CURVE). Let us consider the set of pair of forecasted instants (d_i, d_j) Let us N^2 be the number of possible pairs. Each pair is classified with the following rule :

If the length of confidence interval for instant d_i is lower than the length of confidence interval for instant d_j and the error $\left|\widehat{L}_{d_i,h|(d_i-1)} - L_{d_i}\right|$ for instant d_i is lower than the error $\left|\widehat{L}_{d_j,h|(d_j-1)} - L_{d_j}\right|$ for instant d_i, the pair (i, j) is said *concordant.* On the contrary, if the error of instant i is superior than the error of instant j, the pair is said *discordant.* If length of confidence intervals for the two instants are equal, the pair is said neutral. The highest the proportion of concordant pairs, the best confidence intervals discriminate situations.

Let us define n_c the number of concordant pairs, n_d the number of discordant pairs, and n_n the number of neutral pairs. The measure of discrimination we proposed is : $D = \frac{n_c + n_n/2}{n_n + n_c + n_d}$. The closer this measure is to 1, the best the confidence interval can discriminate different situations. In other words, D is an evaluation of the probability to generate a concordant pair. In case of a neutral pair, there is not a discrimination between situations ; the judgement is neither good nor bad, so we add the quantity 1/2 in D. This is explaining the term $n_n/2$. Let us conclude the description of the measure of discrimination with pointing out that the aim of this criterium is not the error measure but the capacity to discriminate situations.

Table IV shows discrimination measures for the adaptive model and for basic model. Because the width of CIs calculated with basic Model is by definition constant, its measure of discrimination is near 50%. On the contrary, the measure for adaptive model is near 60%. We conclude the adaptive model has more capacity to discriminate situation with a high risk of error from those with a low risk

TABLE IV : DISCRIMINATION MEASURES FOR EACH CASE

Model	Confidence Level (expected)	80%	90%	95%
Adaptive model	7:00	57%	56%	55%
	13:00	62%	62%	61%
	19:00	60%	60%	60%
Basic Model	7:00	47%	50%	52%
	13:00	49%	51%	50%
	19:00	50%	50%	53%

IV. APPLICATIONS

CIs provide a way of quantifying the uncertainty of the forecast. We give here some perspectives of applications for TSO's.

First it could be used to assess as precisely as possible the operating margins. Margins are reserves that are calculated a few days ahead to a few hours ahead face the risk of a loss of production or an overestimation of consumption. Currently, margins are just calculated with a coefficient time the standard deviation of day ahead load forecast error. There is one coefficient by season (winter, summer and interseason). Using a confidence interval with a level to define could assess margins with more accuracy margins, taking into account effects like hour of the day, type of day and the period of year.

The second field of application are security analysis studies ; these studies are done from day-ahead to anticipate real time difficulties. One of the important hypothesis is just the load forecast. We should use one of the quantile calculated with confidence interval to generate extreme demand scenario, anticipating the risk of load under-estimation. Currently, this risk is taken into account by adding a constant value based on expert judgment.

Finally, let us point that others forecasts are derived from day-ahead forecast, for example load prices or power losses. The knowledge of confidence interval should be also used to derive probability law and confidence interval for these quantities.

V. Conclusion

This paper presents a method to obtain CIs for load forecast. It is based on the calculation of empirical quantiles of relative forecast error observed in the past. A classification is made between days for which the load forecast is difficult and those for which it is easier ; this classification is made *a priori*, based on error forecasts knowledge. Next, some CI's skills for the confidence interval are defined evaluated on a test period and show this method is more accurate and useful than a basic method constructed simply with the standard deviation of error.

An improvement of this method could be to find an algorithm that would make the classification automatically. Some tools exist to achieve this aim such as regression trees or logistic regression, but they have to be adapted to the specific problem of interval confidence construction.

CIs provide a way of quantifying the uncertainty of the forecast. For TSOs, application fields are important. It could be used to assess as precisely as possible the operating margins. It could also be used to generate extreme demand scenarios, at a given risk level, for security analysis studies.

The complexity of the problem may even increase in the future with demand response integration initiatives or with the development of distributed production units like photovoltaic modules (e.g. building integrated photovoltaic).

VI. References

Periodicals:

[1] A.P. Alves Da Silva, L.S. Moulin "Confidence Intervals for Neural Network Based Short-Term Load Forecasting " *IEE transactions on power systems*, 2000, vol 15, no4, pp1191-1196.

[2] W. Charytoniuk, J. Niebrzydowski, ; "Confidence interval construction for load forecast", *Electric power systems research,* 1998, vol.48, no2, pp97-103.

[3] P.F. Christoffersen, "Evaluating Interval forecasts", *International Economic Review*, 1998, vol 39, pp 841-62.

[4] J.W Taylor, P.E. Mc Sharry, L.M.M de Menezes, 2006 " A comparison of Univariate Methods for Forecasting Electricity Demand Up to a Day Ahead" *International Journal of Forecasting*, 2006, vol 22, pp 1-16.

Books

[5] D.W. Hosmer, S. Lemeshow, "Applied logistic regression", 2nd edition, John Wiley&Sons, New York, 2000.

[6] M. Tenenhaus, " Méthodes pour décrire, expliquer, et prévoir ", Dunod, 2006.

[7] R. Weron ; "Modeling and forecasting loads and prices : a statistical approach", Wiley Finance, 2006.

Technical Reports:

[8] O. Isengildina-Massa, S.H. Irwin, and D.L. Good, "Empirical Confidence Intervals for WASDE Forecasts of Corn, Soybean, and Wheat Prices*", Marketing and Outlook Research Report 2009-01*, Department of Agricultural and Consumer Economics, University of Illinois at Urbana Champaign, January 2009.

VII. Biography

Benjamin Petiau got an engineer degree from ENSTA (ParisTech) on 2003. He worked at Orange research center, between 2004 and 2006 as a research engineer in statistics applied to economy. Currently, his job at the Power System Department of French TSO Réseau de Transport d'Electricité (RTE), consists in studies on demand forecasting.