To combine or not to combine? Recent trends in electricity price forecasting

Jakub Nowotarski¹
Rafał Weron¹

¹ Department of Operations Research, Wrocław University of Technology, Poland
To combine or not to combine? Recent trends in electricity price forecasting

Jakub Nowotarski*, Rafał Weron*

*Department of Operations Research, Wrocław University of Technology, Wrocław, Poland

Abstract

Essentially everyone agrees nowadays that electricity spot price forecasting is of prime importance to the energy business. A variety of methods and ideas have been tried over the years, with varying degrees of success. Yet, despite this diversity of models, it is impossible to select one single, most reliable approach. We argue here that combining forecasts – also known as averaging forecasts, aggregating experts, committee machines or ensemble averaging – is an idea worth considering. Using publicly available data from the Global Energy Forecasting Competition 2014 and four commonly used time series models, we show that for both point and probabilistic forecasts the quality of predictions can be improved if combined.

Keywords: Electricity price forecasting, Combining forecasts, Ensemble averaging, Aggregating experts, Probabilistic forecasts

1. Introduction

Over the last two decades electricity spot price forecasting has become the core process of an energy company’s planning activities at the operational level [1]. For an electric utility, which under standard (i.e., non-dynamic [2, 3]) tariffs cannot pass costs on to its consumers, the financial consequences of over- or under-contracting and then selling or buying back power in the real-time (i.e., balancing) market are typically so high that they can lead to huge financial losses. At the same time, a generator that is able to forecast the volatile wholesale prices with a reasonable level of accuracy can adjust its bidding strategy and its own production schedule in order to maximize the profits in day-ahead trading.

Essentially everyone agrees nowadays that electricity spot price forecasting is fundamental to the energy business. But it is hard to find information on how much an energy company can actually save, in dollar terms. Since electric load and price forecasts are being used by many departments of an energy company, it is very hard to quantify the benefits of improving them. They will also depend on how large the business is and how accurate are its forecasts now. However, despite these limitations, Hong [4] provides interesting back-of-the-envelope calculations. Based on U.S. data from the last decade, he concludes that savings from a 1% reduction in the mean absolute percentage error (MAPE) for a utility with 1 GW peak load is:

- $300 thousand per year from short-term load forecasting,
- $600 thousand per year from short-term load and price forecasting.

Hence, for a typical medium-size utility with a 5 GW peak load the savings are in millions. Reducing MAPE by 1% may be a too ambitious target for some companies, but even improving the current forecasts by a fraction of a percent will lead to substantial savings.

A variety of methods and ideas have been tried for electricity spot price forecasting, with varying degrees of success. Statistical/econometric approaches (like multiple regressions, autoregressions, (S)AR(J)MA, AR-GARCH, jump-diffusions, factor models and regime-switching models) and computational intelligence techniques (like neural networks, fuzzy techniques and support vector machines) constitute the two main streams of models, both in the academic literature and the business practice itself [1, 5–9]. Yet, despite this diversity of models, it is impossible to select one single, most reliable approach. For instance, Aggarwal et al. [7] compared results from as many as 47 publications and concluded: there is no systematic evidence of out-performance of one model over the other models on a consistent basis. If so, then perhaps we should resort to combining forecasts? Could this give us an edge over competitors?

The main advantage of combining forecasts is not that the best combinations perform better ex-post than the best individual forecasts – as this may not always be the case – but that it is less risky to combine forecasts than to select ex-ante one individual forecasting method [10]. Many combination methods have been proposed over the years, including simple average, Ordinary Least Squares (OLS) averaging, Bayesian methods, and so forth, for reviews see [11, 12]. Simultaneously, approaches known as expert aggregation, committee machines or ensemble averaging that typically involve boosting, bagging or random forests have been developed in the machine learning community [13, 14]. Surprisingly, the researchers from the two groups seem to be unaware of the parallel developments [1]. To a large extent the different terminology used, the different underlyings techniques (neural networks vs. regression models) and the different academic background of the researchers (engineering vs. economics/statistics schools) is the reason for this situation.
Despite its popularity in econometrics in general [15], combining forecasts has not been widely applied in the context of electricity prices. Only very recently, Bordignon et al. [16], Nowotarski et al. [17] and Raviv et al. [18] have provided empirical support for the benefits of combining (point) forecasts to obtain better (point) predictions of electricity spot prices.

Applications of forecast averaging in probabilistic forecasting are even more recent. Nowotarski and Weron [19] have recently introduced a novel method for computing prediction intervals (PI) and dubbed it Quantile Regression Averaging (QRA; see Section 4.2 for a brief account). The method involves applying quantile regression to a pool of point forecasts of individual (i.e. not combined) forecasting models. Using PJM market data they have shown the QRA-implied PI to be more accurate than those obtained for any of the 12 considered individual time series models, both in terms of unconditional and conditional coverage. Using Nord Pool day-ahead prices Nowotarski and Weron [20] have provided even more convincing evidence in favor of the new approach. Also in the Price Track of the Global Energy Forecasting Competition (GEFCom2014) – the largest energy forecasting competition known to date, both by the diversity of competition topics and wide geographic coverage of the participants – the top two winning teams used variants of QRA [21, 22]. This clearly shows the method’s outstanding performance in diverse situations and its practical value.

Using publicly available data from GEFCom2014 and four commonly used time series models, in this article we show that for both point and probabilistic forecasts the quality of predictions can be improved if combined. The remainder of the paper is structured as follows. In Section 2 we describe the dataset and the forecasting setup. Then in Section 3 we briefly present the basic building blocks – the four individual time series models used in this study (AR, ARX, TAR, TARX). In Section 4 we introduce three schemes for averaging point forecasts (SIMPLE, CLS, LAD) and the QRA approach for computing probabilistic forecasts. In Section 5 we analyze the obtained forecasts and in Section 6 we wrap up the results and conclude.

2. The GEFCom2014 dataset

The GEFCom2014 Price Track dataset includes three time series at hourly resolution: locational marginal prices, day-ahead predictions of zonal loads and day-ahead predictions of system loads. During the competition the information set was being extended on a weekly basis to prevent “peeking” into the future. However, now it is available in whole from the www.crowdanalytix.com competition platform. In this paper we only use two subseries – locational marginal prices and day-ahead predictions of zonal loads – and only from the period December 19th, 2011 to December 18th, 2013, see Figure 1. The origin of the data has never been revealed by the organizers, yet given its features it quite likely comes from a region in the Southeastern United States.

We split the dataset into 3 subsets. The first 182 days, December 19th, 2011 – June 17th, 2012 (half a year), are used for calibration of the individual models. When the day-ahead forecasts are made for the 24 hours of June 18th, the window is rolled forward by one day. This procedure is repeated until the predictions for the last day in the sample – December 16th, 2013 – are made. The second period, initially from June 18th to December 17th, 2012 (i.e. 183 days), is used for estimating weights of the forecast averaging models. These, in turn, are used to obtain the day-ahead point (SIMPLE, CLS, LAD) and probabilistic (QRA) forecasts for December 18th, 2012. The window is rolled forward by one day, the weights are recalibrated and combined forecasts are computed for December 19th, 2012. This procedure is repeated until predictions for all 52 weeks in the out-of-sample test period (December 18th, 2012 – December 16th, 2013) are obtained.

3. Individual models

Our choice of individual models includes two autoregressive models (AR and ARX) and two threshold autoregressive models (TAR and TARX). The zonal load is used as the eXogenous variable. We are not very original here as both model classes have been reported to perform well for electricity price forecasting [23–26], also in the forecast averaging context [16, 17].

Following Weron and Misiolek [23], the ARX autoregressive model structure for the natural logarithm of the electricity price is given by the following formula:

\[ y_t = \phi_1 y_{t-24} + \phi_2 y_{t-48} + \phi_3 y_{t-168} + \phi_4 m_t + \psi_1 z_t + d_1 D_{Mon} + d_2 D_{Sat} + d_3 D_{Sun} + \epsilon_t, \quad (1) \]

where the lagged log-prices \( y_{t-24}, y_{t-48} \) and \( y_{t-168} \) account for the autoregressive effects of the previous days (the same hour yesterday, two days ago and one week ago), while \( m_t \) creates the link between bidding and price signals from the entire previous day (it is the minimum of the previous day’s 24 hourly log-prices). The variable \( z_t \) refers to the logarithm of hourly zonal load (actually to a forecast made a day before, see Section 2). The three dummy variables \( D_{Mon}, D_{Sat}, D_{Sun} \) (for Monday, Saturday and Sunday, respectively) – account for the weekly seasonality. Finally, the \( \epsilon_t \)'s are assumed to be independent and identically distributed (i.i.d.) normal variables. Setting \( \psi_1 = 0 \) yields the AR model.

In the threshold models the regime switching between two (or more, in general) autoregressive processes is governed by the value of an observable threshold variable \( v_t \) relative to a chosen threshold level \( T_0 \). The two threshold models used here: TAR and TARX are extensions of AR and ARX, respectively, to two states with \( T_0 = 0 \) and \( v_t \) equal to the difference in mean prices for yesterday and eight days ago, as in [5, 17, 23].

Model parameters are estimated in Matlab by minimizing the in-sample mean squared error and the prediction intervals (PI) are computed by taking a desired quantile of the normal distribution fitted to the in-sample residuals.

4. Combined forecasts

4.1. Point forecasts

In this paper we focus on computing combined spot price forecasts with three averaging schemes – SIMPLE averaging,
least absolute deviation (LAD) regression and constrained least squares regression (CLS). All three schemes have been found to provide accurate and robust results [17]. The first one, SIM- PLE, is the most natural approach to forecast averaging. The idea is to use the arithmetic mean of the (point) forecasts of all individual models. It is highly robust and is widely used in business and economic forecasting [27].

Classical linear regression is another popular, yet not so robust averaging method. In this approach, the individual forecasts are regressors and the corresponding observed spot price is the dependent variable [28]. Nowotarski et al. [17] proposed to replace the ordinary least squares approach with the absolute loss function to yield least absolute deviation (LAD) regression. An advantage of using the absolute loss function is its robustness to electricity price spikes. Indeed, a model that performs well in general, yet significantly underperforms on specific dates, is punished more by the quadratic loss function. As a consequence it leads to a relatively large decrease of this model’s weight, while using the absolute loss function yields a relatively smaller decrease of the weight.

Another variant of OLS regression assumes that the weights in the combined model are constrained to a pre-defined region, hence the name constrained least squares (CLS). Following [18], we restrict them to be nonnegative and add up to one. Note that there is no closed form solution for the LAD and CLS averaging schemes. However, they can be solved using linear and quadratic programming, respectively.

4.2. Probabilistic forecasts

The task gets more challenging when probabilistic forecasts are considered. In particular, applying equal weights to interval forecasts will not ensure the nominal coverage rate, because the mixing of distributions is governed by different rules. The weights have to change with the quantile, and the estimation gets much more complex than for point forecasting [11].

A plausible solution is to apply quantile regression to point forecasts of a set of individual forecasting models, as proposed by Nowotarski and Weron [19]. In their Quantile Regression Averaging (QRA) method the individual point forecasts and the corresponding observations (here: the electricity log-prices, \( y_t \)) are put in a standard quantile regression setting [29]. The quan-
The quantile regression problem can be written as follows:

\[ Q_q(y|X) = X\beta_q, \]  

(2)

where \( Q_q(y|X) \) is the conditional \( q \)-th quantile of the electricity log-price distribution \( y \), \( X \) are the regressors (explanatory variables) and \( \beta_q \) is a vector of parameters for quantile \( q \). The parameters are estimated by minimizing the loss function for a particular \( q \)-th quantile:

\[
\min_{\beta_q} \left\{ \sum_{i=1}^{n} q(1-q)Y_i - X\beta_q \right\} = \min_{\beta_q} \left\{ \sum_{i=1}^{n} \left( q - 1 \right) Y_i - X\beta_q \right\}, \tag{3}
\]

where \( Y_i \) is the actual log-price and \( X_i = [1,\tilde{y}_{1,i},...,\tilde{y}_{m,i}] \) is a vector of point forecasts of \( m \) individual models, see Fig. 2. The choice of the number of individual models can be made arbitrarily (e.g. the best three models, all models, etc.).

If the number of individual models is large, say more than a dozen or two, it may be preferable to first apply a dimension reduction method (like Principal Component Analysis) and then apply QRA to the obtained factors, not the forecasts of the individual models themselves, see Fig. 3. The resulting approach is known as Factor Quantile Regression Averaging (FQRA) and has been originally proposed in [30]. Since here we are working with only four individual models, there is no need to use FQRA.

Finally, note that QRA is a natural extension of point forecast averaging discussed in Section 4.1. In particular, for \( q = 0.5 \) the right hand side of formula (3) under the minimum reduces to \( \frac{1}{2} \sum \left| Y_i - X\beta_q \right| \), which is the function that is minimized in LAD regression [17].

5. Empirical results

We now present the out-of-sample forecasting results for the considered GEFCom2014 dataset. Recall, that we examine day-ahead forecasts of hourly prices for the period December 18, 2012 – December 16, 2013, which is 52 weeks in total.

Forecasts for the considered models are determined the following way: models (as well as model parameters and combination weights) are re-estimated on a daily basis and a forecast for all 24 hours of the next day is determined at the same point in time. Forecasts are first calculated for each of the 4 individual models and then combined according to estimated weights for each of the three point forecasting schemes (SIMPLE, LAD and CLS) and the QRA probabilistic forecasting technique.

5.1. Results for point forecasts

Following [23] and [31], we compare the methods in terms of the Weekly-weighted Mean Absolute Error (WMAE) loss function and evaluate the forecast performance using weekly time intervals, each with \( 24 \times 7 = 168 \) hourly observations. For each week we calculate the WMAE for method \( i \) as:

\[
\text{WMAE}_i = \frac{1}{168} \sum_{h=1}^{168} \left| \hat{Y}_{ih} - Y_h \right|, \tag{4}
\]

where \( Y_h \) is the actual price for hour \( h \) (not the log-price \( y_h \)), \( \hat{Y}_{ih} \) is the predicted price for that hour by model \( i \), and \( \hat{Y}_{168} = \frac{1}{168} \sum_{h=1}^{168} Y_h \) is the mean price for a given week.

The results are summarized in Table 1 in terms of the mean WMAE over all 52 weeks of the out-of-sample test period (WMAE) and the mean deviation from the best model in each week (m.d.f.b.). The latter measure indicates how similar is a model’s performance to the ‘optimal model’ composed of the overall best performing (individual or combined) model in each week.

Out of the three analyzed combination schemes the most accurate is SIMPLE, with \( \bar{\text{WMAE}} = 11.26 \) and m.d.f.b. = 0.63. Yet, clearly all three forecast combination schemes outperform each of the individual models, both in terms of WMAE and m.d.f.b. These results are qualitatively similar to the ones obtained for different datasets and different individual models in [16, 17]. Also in a recent load forecasting study the SIMPLE averaging scheme stands out as one of the best and most robust performers [32].
Table 1: The mean WMAE (i.e., WMAE) over all 52 weeks of the out-of-sample test period and the mean deviation from the best model in each week (m.d.f.b.). The best results in each row are emphasized in bold.

<table>
<thead>
<tr>
<th>Model</th>
<th>WMAE</th>
<th>SIMPLE</th>
<th>CLS</th>
<th>LAD</th>
<th>AR</th>
<th>TAR</th>
<th>ARX</th>
<th>TARX</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>11.26</td>
<td>11.33</td>
<td>11.48</td>
<td>11.71</td>
<td>12.00</td>
<td>11.68</td>
<td>11.64</td>
</tr>
<tr>
<td>m.d.f.b.</td>
<td>0.63</td>
<td>0.69</td>
<td>0.84</td>
<td>1.08</td>
<td>1.37</td>
<td>1.04</td>
<td>1.00</td>
<td></td>
</tr>
</tbody>
</table>

5.2. Results for probabilistic forecasts

In probabilistic forecasting our preference is to choose the model that yields the narrowest prediction intervals (PI) out of those that provide (unconditional) coverage close to nominal. The score function proposed by Winkler [33], usually referred to as the Winkler score or interval score, allows to jointly assess the unconditional coverage and interval width. For a central \((1 - \alpha) \cdot 100\%\) prediction interval, it is defined as:

\[
W_t = \begin{cases} 
\delta_t & \text{for } Y_t \in [L_t, U_t], \\
\delta_t + \frac{2}{\alpha}(L_t - Y_t) & \text{for } Y_t < L_t, \\
\delta_t + \frac{2}{\alpha}(Y_t - U_t) & \text{for } Y_t > U_t,
\end{cases}
\]

(5)

where \(L_t\) and \(U_t\) are respectively the lower and upper bounds of the PI, \(\delta_t = U_t - L_t\) is the interval width and \(Y_t\) is the actual price. The Winkler score gives a penalty if an observation (the actual price) lies outside the constructed interval and rewards a forecaster for a narrow PI; naturally the lower the score the better the PI. The results presented in Table 2 confirm that QRA leads to better 50% and 90% PI than the two benchmarks – ARX and TARX.

The Winkler score can be used to compare the different methods, but cannot be used to answer the question of whether the PI obtained are statistically sound. We now apply the approach of Christoffersen [34] to test the conditional coverage of the constructed PI. The idea behind the test is to operate on variables that indicate whether the actual price lies in the pre-constructed prediction interval or not. The conditional coverage (CC) test is a sum of the unconditional coverage (UC) and independence tests. The UC test compares the nominal coverage of the model to the true coverage, and is also known in the risk management (Value-at-Risk backtesting) literature as the Kupiec [35] test. The independence test checks that the PI violations do not cluster. All three tests are carried out in the likelihood ratio (LR) framework and the CC test is distributed asymptotically as \(\chi^2(2)\).

The conditional coverage LR statistics are plotted in Fig. 4. We follow the approach of [19, 23, 30] and conduct the tests separately for each hour, since the forecasts for consecutive hours are correlated by construction – the 24 hourly day-ahead predictions are calculated at the same time and using the same information set. As can be seen in Fig. 4, QRA outperforms both benchmarks (ARX and TARX), for a vast majority of hours and both levels (50% and 90%). In particular, for the 50% intervals at the 5% (1%) significance level, QRA passed the conditional coverage test 16 (20) times, while ARX and TARX were successful respectively for only 3 (7) and 2 (3) hours. For the 90% intervals the results are not as good. But still, at the 1% level QRA passed the test 7 times, while TARX only once and ARX not even a single time.

Table 2: The Winkler score for the 50% and 90% two-sided day-ahead prediction intervals (PI), as defined by Eq. (5). The best results in each row are emphasized in bold.

<table>
<thead>
<tr>
<th>Model</th>
<th>50% PI</th>
<th>90% PI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARX</td>
<td>24.63</td>
<td>56.60</td>
</tr>
<tr>
<td>TARX</td>
<td>24.53</td>
<td>54.16</td>
</tr>
<tr>
<td>QRA</td>
<td>24.13</td>
<td>53.07</td>
</tr>
</tbody>
</table>
6. Conclusions

In this paper we have examined possible accuracy gains from combining forecasts of four individual time series models (AR, ARX, TAR, TARX) for computing point and probabilistic forecasts of electricity spot prices. Overall, our findings support the additional benefits of combining forecasts. All three combination schemes (SIMPLE, LAD, CLS) outperform each of the individual models in point forecasts, both in terms of WMAE and m.d.f.b. Also when probabilistic forecasts are considered, the QRA beats the two benchmarks (ARX and TARX).

We would like to emphasize two things. Firstly, such a good forecasting performance of the combination approaches does not always have to be the case. In general, it depends on the quality of the building blocks. In particular, if one individual model outperforms the other on a continuous basis, there may be no advantage in combining at all. Secondly, we should remember that the main advantage of combining forecasts is not that the best combinations perform better ex-post than the best individual forecasts – as this may not always be the case – but that it is less risky to combine forecasts than to select ex-ante one individual forecasting method.

Acknowledgments

This work was partially supported by the National Science Center (NCN, Poland) through grants 2013/11/N/HS4/03649 and 2015/17/B/HS4/00334.

References

To combine or not to combine? Recent trends in electricity price forecasting by Jakub Nowotarski and Rafał Weron