UNIVERSIDAD CARLOS III DE MADRID working papers

UC3M Working Papers
Statistics and Econometrics
17-01
ISSN 2387-0303
Enero 2017

Departamento de Estadística
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (34) 91624-98-48

ELECTRICITY PRICES FORECASTING BY AVERAGING DYNAMIC FACTOR MODELS

Andrés M. Alonso[a,b], Guadalupe Bastos[a], Carolina García-Martos[c]

Abstract: In the context of the liberalization of electricity markets, forecasting prices is essential. With this aim, research has evolved to model the particularities of electricity prices.
In particular, Dynamic Factor Models have been quite successful in the task, both in the short and long run. However, specifying a single model for the unobserved factors is difficult, and it can not be guaranteed that such a model exists. In this paper, Model Averaging is employed to overcome this difficulty, with the expectation that electricity prices would be better forecast by acombination of models for the factors than by a single model. Although our procedure is applicable in other markets, it is illustrated with applications to forecasting spot prices of the Iberian Market, MIBEL (The Iberian Electricity Market) and the Italian Market. Three combinations of forecasts are successful in providing improved results for alternative forecasting horizons.

Keywords: *Dimensionality reduction; Electricity Prices; Bayesian Model Averaging; Forecast Combination.*

[a]Department of Statistics, Universidad Carlos III de Madrid.
[b]Instituto Flores de Lemus, Universidad Carlos III de Madrid.
[c]Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid.

Acknowledgements: A.M. Alonso acknowledges support of the Spanish Ministry of Economy and Competitiveness, research projects ECO2012-38442, and ECO2015-66593. Carolina García-Martos acknowledges financial support from project DPI2011-23500, Spanish Ministry of Economy and Competitiveness.
The authors would like to extend their appreciation to Professor Michael Wiper for his assistance and corrections regarding the proper use of English in this document.

Working Paper UC3M

Electricity Prices Forecasting by Averaging Dynamic Factor Models

Andrés M. Alonso [1,2], Guadalupe Bastos [1]* and Carolina García-Martos [3]

[1] Department of Statistics, Universidad Carlos III de Madrid, Getafe 28903, Madrid, Spain
[2] Instituto Flores de Lemus, Universidad Carlos III de Madrid, Getafe 28903, Madrid, Spain
[3] Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid, Madrid 28040, Spain
* Correspondence: guadalupe.bastos@alumnos.uc3m.es

Abstract: In the context of the liberalization of electricity markets, forecasting prices is essential. With this aim, research has evolved to model the particularities of electricity prices.
In particular, Dynamic Factor Models have been quite successful in the task, both in the short and long run. However, specifying a single model for the unobserved factors is difficult, and it can not be guaranteed that such a model exists. In this paper, Model Averaging is employed to overcome this difficulty, with the expectation that electricity prices would be better forecast by a combination of models for the factors than by a single model. Although our procedure is applicable in other markets, it is illustrated with applications to forecasting spot prices of the Iberian Market, MIBEL (The Iberian Electricity Market) and the Italian Market. Three combinations of forecasts are successful in providing improved results for alternative forecasting horizons.

Keywords: Dimensionality reduction; Electricity Prices; Bayesian Model Averaging; Forecast Combination.

MSC: 62M10, 62P30, 62M20

---

1. Introduction

Nowadays, electricity trading is liberalized in most countries of the Western world. Due to the particular characteristics of supply and demand, prediction of electricity prices in this context is complex. Notwithstanding the difficulties, forecasts are necessary for several reasons:

- this is a strategic sector of the economy,
- there are financial implications due to the trading of forwards and options,
- forecasts help optimize and plan consumption and production.

As with other commodities, there are various ways to operate in this market (see Weron, 2014, for a detailed market description and a thorough literature review). We focus on prices that result from a *pool* in which there is a central auction. In this *pool*, prices could be settled for each hour of the day, or every half hour, depending on the market.

In the first case, the 24 hourly prices for day t are cleared at the same instant in day $t-1$, with the same common information for all the hours. Therefore, for each day, a 24-dimensional vector is generated $(p_{1,t}, p_{2,t}, \ldots, p_{24,t})$; where $p_{hour,t}$ represents the price of $hour = 1, 2, \ldots, 24$ at day t. Consequently, prices can be presented in a $T \times 24$ dimensional matrix, where T is the number of days in the sample, and modeling should be multivariate (as in Huisman et al., 2007; Panagiotelis and Smith, 2008; García-Martos et al., 2007; Alonso et al., 2011).

In several fields, there has been an increasing interest in the development of methodology to deal with multivariate time series or a high dimensional vector of series like the ones in electricity markets. By the end of the 1970s, Sargent and Sims (1977) (these authors presented a factor model for stationary time series vectors) and Geweke (1977) were the first to propose a Dynamic Factor

Model. Later, Lee and Carter (1992) contributed by extending the idea of Principal Components to the dynamic case. More recently, dimensionality reduction techniques have gained popularity, in particular since the work by Stock and Watson (2002). For example, Peña and Poncela (2004) and Peña and Poncela (2006) extended Sargent and Sims (1977)'s model for the non-stationary case.

Regarding applications in electricity markets, García-Martos et al. (2012) extended Lee and Carter (1992) and Peña and Box (1987) to prices with seasonality. Working with data for the Iberian market for 2007-2009, they propose extracting common factors from the 24-dimensional price vector, and modeling such factors as univariate seasonal AutoRegressive Integrated Moving Average (ARIMA) processes. Another example is Alonso et al. (2011), who propose a technique called Seasonal Dynamic Factor Analysis (SeaDFA), which involves the estimation of a Vector AutoRegressive Integrated Moving Average (VARIMA) model for unobserved common factors having seasonal patterns. The work in Maciejowska and Weron (2015) also uses a Factor Model, including hours and locations.

In an independent path, Forecast Combination or Model Averaging has been developed as a technique to take advantage of the availability of alternative forecasting approaches. This methodology consists of weighting a set of forecasts corresponding to alternative models, and combining them to obtain a single forecast. In this way, model selection uncertainty is incorporated. According to Clemen (1989), 'the idea of combining forecasts implicitly assumed that one could not identify the underlying process, but that different forecasting models were able to capture different aspects of the information available for prediction'. Other justifications for model averaging are: doubts of the existence of a 'best model' (Sánchez, 2006), 'portfolio diversification', a better adaptation to structural breaks, or to average out omitted variables bias (Bjørnland et al., 2010).

Applications of model averaging in electricity markets are given by Bordignon et al. (2013), for the British Market, and Nowotarskia et al. (2014), for European and USA markets. Furthermore, Raviv et al. (2015) obtain forecasts for the daily average price employing dimensionality reduction techniques as well as Forecast Combination of several models for hourly prices. Other references are Monteiro et al. (2015), who use averaging to obtain wind speed, solar irradiation, and temperature forecasts, which are then employed to estimate prices; and García-Martos et al. (2015), who forecast hourly electricity prices for the Spanish market by weighting seasonal ARIMA (with exogenous variables) and seasonal Dynamic Factor Models of similar performance.

In spite of its advantages, a major drawback of dimensionality reduction techniques is the uncertainty concerning the 'correct' model: how many factors to include, and what models they follow. The literature is not definite in regard to the best technique for estimating the number of underlying factors that would contain enough information to make accurate predictions, and it should be considered that, as the number of factors included increases, so does estimation complexity and computational burden. As previously indicated, there is not either a unique model for the factors that outperforms all other models, in all circumstances (Weron, 2014).

In this work it is assumed that the major decisions attached to forecasting by using dimensionality reduction techniques may be resolved in a less arbitrary way if Forecast Combination is included. In order to follow this line of thought, alternative models, including different numbers of common factors, are estimated. Forecasts for prices are obtained by transforming the factors' forecasts back to the data units, according to the relations established in the dimensionality technique employed. Subsequently, Forecast Combination approaches are used to weight each of the forecasts obtained, and thus provide a single prediction.

Summing up, factor models extract information *ex ante* (before any forecast is obtained) while Forecast Combination works *ex post* (after forecasts are available). The contribution of this work is to amalgamate both techniques. A reduced number of latent unobserved variables is estimated, and their forecasts are combined in order to obtain a single prediction.

We apply these techniques to one-day to two-month-ahead electricity prices for the Iberian spot Market for a period of five years (2008-2012); and for the Italian spot Market for a period of three

and a half years (mid 2009-2012). Several ARIMA specifications[1] are estimated for the factors, and used to obtain forecasts of the prices for each hour, which makes the task computationally intensive. Next, these forecasts are combined. We study alternative ways to combine forecasts because their performance may vary depending on the data-set. The predictions concern mainly the short and medium-term (one and two months), but a one-year extension is presented to illustrate the potential accomplishments in long-term forecasting.

The rest of the paper is organized as follows. Fundamentals containing a mathematical description of the proposed methodology are presented in Section 2, which includes definitions on Dynamic Factor Models, classical techniques for Forecast Combination, and Bayesian Model Averaging. Section 3 describes the methodology for this paper. In section 4, we present the results of the empirical applications. This section is divided into sub-sections presenting the data, an Analysis of Variance (ANOVA) comparing specifications, and forecasting results. Finally, section 5 concludes with remarks, limitations and possible extensions.

2. Fundamentals

An outline of the methodology used in this proposal is presented below, as are the drawbacks of other approaches, which we attempt to resolve.

2.1. Dynamic Factor Model

Dynamic Factor Models (DFM) are a widely applied dimensionality reduction technique. It is employed when the researcher believes there are fundamental factors driving several variables in a data-set. These factors, like the variables, evolve through time, and allow to obtain information about the larger data-set with a simpler model. The explanation here follows García-Martos et al. (2012). As there, once the common factors are obtained, univariate seasonal ARIMA models are fitted to them. The forecasts of these models are then combined to obtain one improved forecast.

Let y_t be an N-dimensional observed time series vector, generated by an R-dimensional vector of unobserved common factors $R \ll N$. In the Iberian and Italian electricity markets $N = 24$, and the matrix of observed series has as many rows as days are considered in the historic data-set. As in Lee and Carter (1992), it is assumed that vector y_t can be written as a linear combination of the unobserved common factors F_t, plus a vector of specific components or factors ε_t:

$$y_t = \Omega F_t + \varepsilon_t, \tag{1}$$

where Ω is an $N \times R$ matrix of loads relating the set of R common unobserved factors with the vector of observed series y_t (the vector of the 24 hourly prices for our application), and ε_t is an N-dimensional vector of specific components.

To estimate the factors F_t, singular value decomposition (SVD) is used (as in Lee and Carter, 1992) for the covariance of the 24 dimensional vector of centred prices (García-Martos et al., 2012). This consists in calculating the eigevalues, and their associated eigenvectors, for the sample covariance matrix, and thereupon calculating the matrix of common factors, f, as a linear combination of the time series: $f_{T\times R} = Y_{T\times N}\hat{\Omega}_{N\times R}$.

The common factors F_t may be non-stationary, including regular or seasonal unit roots in addition to auto-regressive and moving average (regular and seasonal) components. These

[1] For each one of the common factors included in the analysis 36 choices of parameters are available: $p = 1,2,3$, $d = 0$, $q = 1,2,3$, $P = 0,1$, $D = 1$, $Q = 0,1$, $s = 7$. These pre-defined models are all automatically estimated with the software TRAMO, by its Matlab interface, intervening outliers.

ARIMA$(p,d,q) \times (P,D,Q)s$ models are used to obtain factors' forecasts, and from them prices' forecasts. For instance, the i-th factor at date t, F_{it}, would be modeled by

$$(1-L)^d(1-L^s)^D\phi_i(L)\Phi_i(L^s)F_{it} = c_i + \theta_i(L)\Theta_i(L^s)\eta_{it}, \tag{2}$$

where $i = 1,2,\ldots,R$ is the i-th factor, $\phi_i(L) = (1-\phi_{i1}L-\phi_{i2}L^2-\ldots-\phi_{ip_i}L^{p_i})$, $\Phi_i(L^s) = (1-\Phi_{i1}L^s-\Phi_{i2}L^{2s}-\ldots-\Phi_{iP_i}L^{P_is})$, $\theta_i(L) = (1-\theta_{i1}L-\theta_{i2}L^2-\ldots-\theta_{iq_i}L^{q_i})$, and $\Theta_i(L^s) = (1-\Theta_{i1}L-\Theta_{i2}L^{2s}-\ldots-\Theta_{iQ_i}L^{Q_is})$ are polynomials, L is the lag operator such that $Ly_t = y_{t-1}$. The roots of $|\phi_i(L)| = 0$, $|\Phi_i(L^s)| = 0$, $|\theta_i(L)| = 0$, $|\Theta_i(L^s)| = 0$, satisfy the usual stationarity and invertibility conditions, and $\eta_{it}\sim N(0,W_i)$ are uncorrelated $E(\eta_{it}\eta'_{it-h}) = 0, h \neq 0$. It is also assumed that the error term of the common factors η_{it} is uncorrelated with the specific components $E(\eta_{it}\varepsilon'_{t-h}) = 0, \forall h$. c_i is the constant of the model for the common factors, and its inclusion in the common factors model (2) can be particularly relevant to calculate long-term forecasts in the non-stationary case, which is the case of electricity prices. Furthermore, in this work the specific components are assumed to be independent and have no dynamic structure along them (e.g. Peña and Poncela, 2006).

It should be noticed that we work in two consecutive steps: firstly we estimate the factors f, and secondly we estimate the ARIMA models like (2) ($\hat{\phi}, \hat{\Phi}, \hat{\theta}, \hat{\Theta}$ are estimated for each common factor f_i). Morever, the estimation of the first factor is the same when $R = 1$ or $R > 1$, which is natural consequence of the SVD procedure. Nevertheless, the selection of R affects the forecasting errors for the series. The more factors are included (greater r), the greater the variability of the data explained by the model. The cost of incorporating more factors is an increase in the number of parameters to estimate.

To summarize, a key stage when estimating this kind of models is the selection of the number of common factors, R, as well as the model they follow, which implies selecting the orders: p,d,q,P,D,Q. r could be obtained using existing tests such as the ones proposed in Peña and Poncela (2006) or Bai and Ng (2002), and could also be selected such that diagnostic checking results[2] are reasonable (Alonso et al., 2011). However, alternative values could satisfy these criteria. Because selecting one value for R and the other parameters will likely not render the best results in every scenario, we will instead keep the alternatives and combine their forecasts. Forecast Combination is presented in the following Subsection 2.2.

2.2. Forecast Combination

Empirically, the improvements of using Forecast Combination instead of a "best" model have been shown for different types of models (for instance see Poncela et al., 2011; Kuzin et al., 2012; Martínez-Álvarez et al., 2015), and in various research areas (Clemen, 1989; Stock and Watson, 2004). However, Weron (2014) points out that Forecast Combination techniques have not been fully exploited for electricity prices.

In general, we can think of the combination equation as follows:

$$\hat{y}^C_{t+h|t} = \sum_{i=1}^{K} w^{(h)}_{t,i}\hat{y}^{(i)}_{t+h|t}, \tag{3}$$

where $w^{(h)}_{t,i}$ is the *i-th* model weight at time t for the forecast horizon h, K the total number of models considered, and $\hat{y}^{(i)}_{t+h|t}$ the forecast obtained with the *i-th* model for the forecast horizon h.

Combinations will vary depending on the weights they use, and the set of models they include. There are classical and Bayesian techniques. In the next subsections, we briefly summarize the literature on both classical approach and an approximation to Bayesian combination, mentioning

---

[2] Specific factors and errors of the observation equation must be uncorrelated between them, and specific factors without any cross correlation.

their drawbacks and advantages. This will help us justify our methodological proposal presented in Section 3.

2.2.1. Classical techniques for Forecast Combination

One easy way to obtain Forecast Combination is the simple average, in which all alternative forecasts are given the same weight. This approach often works very well in comparison with more complex ones. One possible reason is that 'complicated combining methods pursuing "optimal" behavior often lead to unstable weights and the combined forecast even performs significantly worse than the individual forecasts' (Yang, 2004). Alternatively, a simple combination method outperforming more complex ones might be explained by a larger variability of the latter (Yang, 2004). In this regard, Bjørnland et al. (2010) advice to use a simple average when the alternative models to combine have similar forecast error variance.

A different approach to assign weights consists of estimating weights that minimize a loss function with the forecast error of the models to combine as explanatory variables (Elliott et al., 2006).

A further option is a combination using only the d_m best models. Possible advantages of this approach are: to reduce the variability of the combination (Yang, 2004), and to avoid under-weighting independent information when the models are correlated (Bjørnland et al., 2010). The set d_m could change through time depending on the most recent performance of the models (Bjørnland et al., 2010)[3], or it could be fixed (Swanson and Zeng, 2001).

Another way to combine forecasts would be to employ the median prediction (Kuzin et al., 2012). Alternatively, some authors employ a combination regression of the form

$$y_{t+h} = \alpha_0 + \sum_{i=1}^{P} \alpha_i p_{i,t} + \epsilon_{t+1},$$

where y_{t+h} is the forecast resulting from the combination, and $p_{i,t}$ are the predictions of the alternative models. Swanson and Zeng (2001) use the Bayesian Information Criterion (BIC)(Schwarz, 1978) or the Akaike Information Criterion (AIC) to select the best combination. There are also some drawbacks to this regression based approach. Swanson and Zeng (2001) indicate collinearity in the competing forecasts, and over-fitting due to outliers; Wright (2008) adds that while in-sample fit is improved, out-of-sample prediction tends to be worse than using the average to combine.

Even using complex combinations, empirical findings in Swanson and Zeng (2001) suggest that, in some cases, the difference between alternative combination methods is not significant, a result that will also be obtained at points in this work.

2.2.2. Bayesian techniques for Forecast Combination (BMA)

With this approach, the predictive distribution of a new observation is obtained by averaging with different weights the predictive distribution of each model considered. The idea was initially introduced by Leamer (1978) and allows to incorporate the uncertainty regarding the variety of available models (Leamer, 1978). It has been applied in statistics (Raftery, 1995; Raftery et al., 1997; Chipman et al., 2001) and econometrics (Koop and Potter, 2003; Cremers, 2002).

According to Wright (2008), an advantage of Bayesian Model Averaging (BMA) is that 'One does not have to be a subjectivist Bayesian to believe in the usefulness of BMA, or of Bayesian shrinkage techniques more generally. A frequentist econometrician can interpret these methods as pragmatic devices that may be useful for out-of-sample forecasting in the face of model and parameter uncertainty'.

---

[3] These authors evaluate model performance based on the sum squared errors. The results they obtain with a time-varying subgroup of models outperform those of the simple average of all the models.

As Wright (2008) explains, the procedure takes under consideration a large number of alternative models' forecasts, assuming one of them is the 'true' data-generating model; however, the researcher is unaware of which one is this. A prior regarding which model is the correct one is set, and then *a posteriori* probabilities of the different models being the true one are obtained to weight the predictions.

Alternative models' weights can be time evolving. For instance, Billio et al. (2011) work with weights that change depending on the predictive densities past performance and learning mechanisms.

Following Wright (2008): let K be the total number of models $M_1, \ldots, M_K$. The i-th model is related to the vector of parameters θ_i. The researcher has *a priori* knowledge of the probability that the i-th model is the true one, $p(M_i)$. Then the data, D, is observed and the probability is updated by calculating the *a posteriori* probability that model i-th is the true one:

$$p(M_i|D) = \frac{p(D|M_i)p(M_i)}{\sum_{i=1}^{K} p(D|M_j)p(M_j)}, \tag{4}$$

where $p(D|M_i) = \int p(D|\theta_i, M_i)p(\theta_i|M_i)d\theta_i$ is the marginal likelihood of the i-th model, $p(\theta_i|M_i)$ is the *a priori* density of that model parameters vector, and $p(D|\theta_i, M_i)$ is the likelihood. Inference about a 'future' quantity Δ is based on

$$p(\Delta|D) = \sum_{i=1}^{K} p(\Delta|D, M_i)p(M_i|D). \tag{5}$$

In particular, the mean of this posterior distribution is used as forecast. This procedure minimizes the Mean Squared Forecast Error (MSFE). It is only necessary to specify the set of models, their priors $p(M_i)$, and the parameters' priors $p(\theta_i|M_i)$.

A disadvantage of this approach though, is that the conditional probabilities are, in general, unknown. Therefore, they should be estimated from the data, which could mean that any benefits of Forecast Combination are lost.

Often, all models will have equal *a priori* probabilities, i.e. $p(M_i) = 1/K$. In this case, as Raftery (1995) indicates, the posterior probability $p(D|M_i)$ is proportional to $\exp(-(1/2)BIC_i)$. Therefore, expression (4) can be written as follows,

$$p(M_i|D) \approx \frac{\exp(-(1/2)BIC_i)}{\sum_{i=1}^{K} \exp(-(1/2)BIC_i)}. \tag{6}$$

Expression (6) is easy to calculate and no prior densities need to be set (Raftery, 1995). In this paper, one of the Forecast Combinations will use weights obtained as indicated in expression (6).

Notice that the selection of equal *a priori* probabilities is motivated by the approach of using non-informative *a priori* probabilities. However, other *a priori* probabilities can be considered and, in such cases, expression (6) would be:

$$p(M_i|D) \approx \frac{\exp(-(1/2)BIC_i)p(M_i)}{\sum_{i=1}^{K} \exp(-(1/2)BIC_i)p(M_i)}. \tag{7}$$

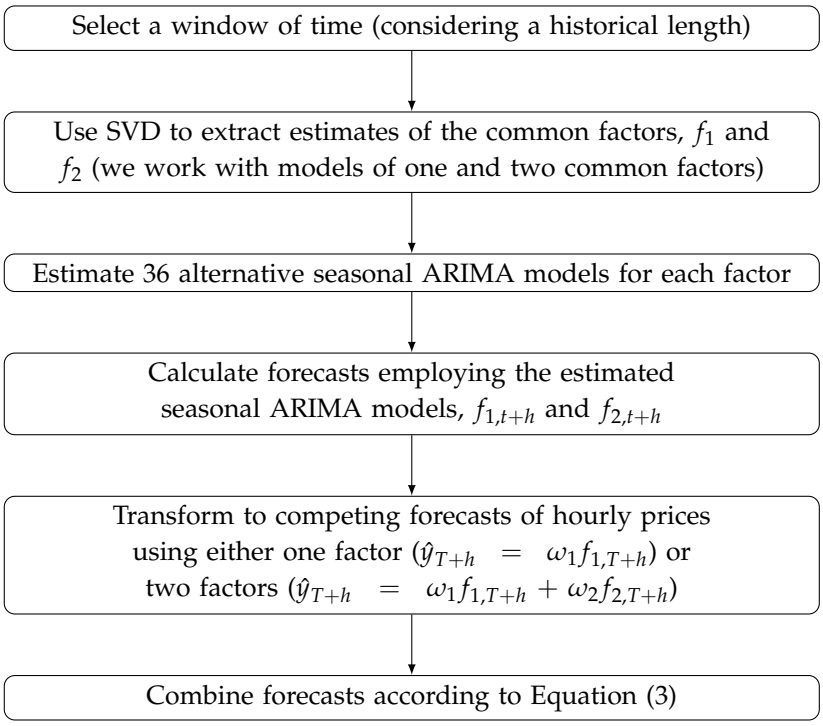
In this paper the goal is to derive some feasible and reasonable weights, not to estimate conditional probabilities. Of course, it is to be expected that clever *a priori* probabilities produce better weights in the sense of better forecast performance.

3. Methodology

Taking into account the limitations of existent approaches in dimensionality reduction, most importantly the issue of selecting a number of common underlying factors r, as well deciding for a 'best' model for them; and given the advantages of Forecast Combination revisited in the previous sections, our methodological proposal consists of averaging the forecasts of alternative models for each factor.

This allows to capture the factors underlying the behavior of large data-sets, avoiding the risk of committing to a particularly 'bad' specification for them. That is why we consider that this approach improves previously mentioned solutions to open problems described along sections 1 and 2.

The complete prediction procedure can be summarized in the following steps, repeated for each window of time in the data-set. Notice that each window of time provides a historical data-set as well as out of sample data with which the forecasts will be compared.



For each window of time, the factors underlying the data are estimated by means of SVD, as explained in Section 2.1. There are as many common factors as time series in the data-set, N. However, the purpose of applying dimensionality reduction techniques is to be able to describe the data by means of a much smaller number of variables, thus $R << N$. There are many criteria for estimating the value R that would best represent the underlying trends in the data. In this regard, a contribution of this work is that, instead of committing to one of them, the possibility of estimating several models is explored. For this reason, at least two settings are estimated: on the one hand $r = 1$, which means that only the underlying factor most representative of the data variability is used to forecast; and on the other hand $r = 2$, which means that the first and second most important underlying factors are estimated and employed to obtain forecasts. Based on the percentage of the total variability explained by the common factors, having up to $r = 2$ in the case of the Iberian data corresponds to explaining about 80% of the total variability. However, for the Italian data we need $r = 3$ to achieve a similar result in terms of the percent of variability explained. Therefore, we incorporate the option of having up to three common factors in the models, and the steps described should accommodate a third factor f_3 for this data-set.

As indicated in the flowchart, the next step consists of estimating models for the factors. The literature review performed in this work reveals that it is difficult, if not impossible, to find a model

that by all criteria would outperform all others. Even more, a good fit does not guarantee an accurate forecasting performance. To overcome these difficulties, our proposal consists of fitting 36 ARIMA specifications for each estimated factor, in lieu of selecting a 'best' set of parameters. These specifications result from the following parameters: $p = \{1,2,3\}$, $d = \{0\}$, $q = \{1,2,3\}$, $P = \{0,1\}$, $D = \{1\}$, $Q = \{0,1\}$, $s = 7$. Additional values of the parameters (for example, $p > 3$) are excluded because they increase the computational burden but do not provide a relevant improvement in results.

After forecasts are estimated for all the options of factors (either one or two) and ARIMA models, they are transformed to forecasts for the original variables, by means of a multiplication by the matrix of weights following expression (1). This will render many forecasts for the data, which will be combined to present with a single forecast for each variable of the original data-set.

3.1. Forecast Combinations and Accuracy Metrics

We consider five alternative combinations (2 to 6 below), and compare them to a benchmark (1 in the next enumeration):

1. Forecast resulting from the benchmark model ('BIC-selected model' for future reference). This is the best model according to the BIC (has the lowest BIC). Selecting only one model is equivalent to assigning it a weight $w_i = 1$ in (3), and $w_i = 0$ for all other models. The superscript (h) has been eliminated from expression (3) because weights will not be adaptive to the forecasting horizons, and subscript t has also been omitted to avoid confusion with time-varying weights.
2. Forecast calculated as the median of the forecasts of all the models ('median-based combination'). This is also a case of weights $w_i^{(h)} = 1$, for the model with the median forecast, and $w_i^{(h)} = 0$ for all other models.
3. Forecast equal to the mean of all forecasts ('mean-based combination'). In this case, expression (3)'s weights are all equal $w_i = 1/K$, where K is the total number of models in the analysis.
4. Forecast obtained using BIC-based weights as in expression (6) ('BIC-based combination'). This approach involves equal *a priori* probabilities. Other sensible sets of *a priori* probabilities were considered, and similar results were obtained.
5. Forecast obtained with BIC-based weights for the top 50% models ('BIC-50% combination'). In other words, half of the models are included according to their BIC criterion $w_i = p(M_i|D)$ of expression (6), and for the half that has the largest BIC values, $w_i = 0$. Let us recall that the BIC evaluates the fit of the model, not how accurate it is when used to forecast.
6. Forecast calculated as the mean of the forecasts of the top 50% models ('mean BIC-based combination'). Only half of the models are included (the 'best' half models depending on their BIC), and the Forecast Combination is simply their average. In other words, the 50% models with the lowest BIC are assigned weights $w_i = 2/K$, and the 50% models with the greatest BIC are assigned weights $w_i = 0$.

In order to evaluate forecasts and assess the most appropriate combination, we need to define a forecasting accuracy metric. We can evaluate the forecasts' accuracy by means of several alternative metrics, see Conejo et al. (2005); Hyndman and Koehler (2006); Weron (2014) for a detailed review. Some of them are the relative forecast error, and the Mean (and Median) Average Percentage Error (MAPE). However, these measures are not valid when the data have negative and/or positive, but close to zero, values (Hyndman and Koehler, 2006), a frequent occurrence for many electricity markets (Bello et al., 2016; Monteiro et al., 2015, deal with the issue of forecasting extreme prices in the Spanish electricity market).

Therefore, we use the Mean Absolute Error (MAE) and Median Absolute Error (MedAE). They can be obtained as follows,

$$MAE_\tau^i = \frac{1}{m} \sum_{z=\tau+1}^{\tau+m} \left(\frac{1}{24} \sum_{h=1}^{24} |(y_{h,z} - \hat{y}_{h,z}^i)| \right) \tag{8}$$

and

$$MedAE^i_\tau = \frac{1}{m}\sum_{z=\tau+1}^{\tau+m}(median(|y_{h,z} - \hat{y}^i_{h,z}|)), \tag{9}$$

where m is the number of days in the out-of-sample period, and τ is the last observation of the rolling window employed to estimate the model used to compute the forecasts.

Additionally, in order to simplify the comparison between the benchmark and Forecast Combinations the Relative MAE (RelMAE) will be computed. Following Hyndman and Koehler (2006), we calculate it as follows,

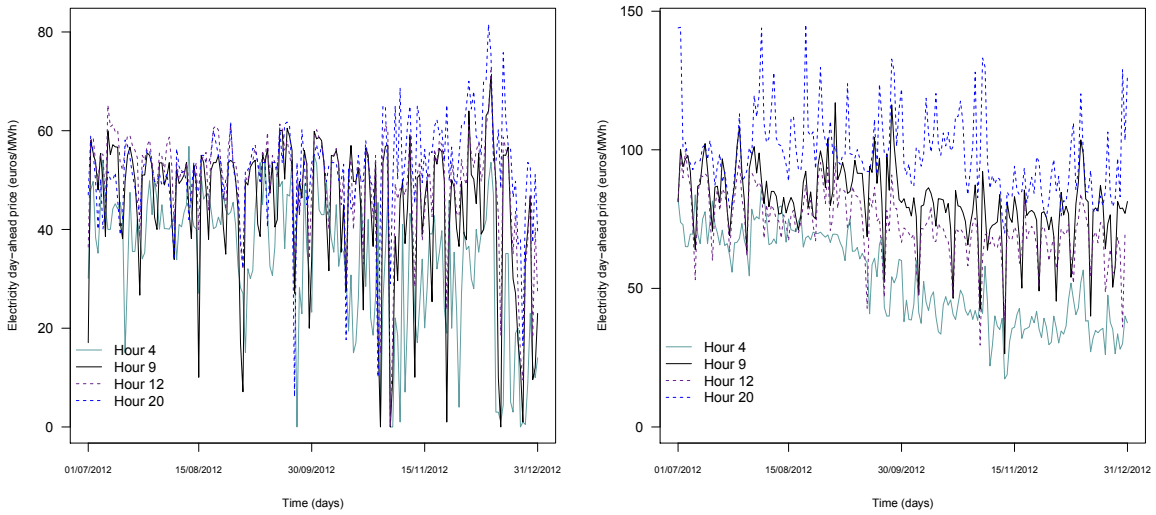
$$RelMAE^i_\tau = MAE^i_\tau / MAE^b_\tau, \tag{10}$$

where b indicates the benchmark model (BIC-selected model). As indicated in Hyndman and Koehler (2006), whenever $RelMAE^i_\tau < 1$ the forecast provided by the i-th combination is better than the one provided by the benchmark, and the opposite happens when $RelMAE^i_\tau > 1$.

4. Results

4.1. Data

We study two data-sets of electricity spot prices, one for the Iberian market, which includes Spain and Portugal (July 2006 - December 2012), and the other one for the Italian market (January 2008 - December 2012). To illustrate the behavior of these prices, a few representative time series are plotted. To the left, Figure 1a, corresponds to the Iberian market, while to the right, Figure 1b, corresponds to the Italian market. Both figures present hours 4, 9, 12, and 20, for the last six months of 2012. In each figure there is a common pattern in the evolution of the hourly series, which is what the common factors attempt to capture.



(a) Iberian market.

(b) Italian market.

Figure 1. Electricity day-ahead prices for four representative hours during the last semester of 2012.

4.2. ANOVA for Comparison of Alternatives for Modeling

We rely on Design of Experiments (DOE) techniques to assess which is the forecasting methodology that produces the smallest error, measured by MAE, in the forecasts of electricity prices. We consider the following factors:

- *Logarithm*: this factor has two levels, $Logarithm = \{No, Yes\}$. $Logarithm = No$ when we use the prices in the same way they are reported (€ per MWh). $Logarithm = Yes$ means that we work with $ln(prices)$. Taking logarithm has the effect of producing time series with less volatility.
- *Historical Length*: this factor indicates how long is the dataset employed in each rolling window. It has two levels, $Historical\ Length = \{308\ days,\ 548\ days\}$. $Historical\ Length = 308\ days$ indicates that the common factors are extracted from series of prices with an extension of 44 weeks (García-Martos et al., 2012). $Historical\ Length = 548\ days$ involves employing time series that extend for 1.5 years, supporting the well known idea that the estimation of common factors benefits from extensive data.
- *Moving Average (MA)*: this factor has two levels, $MA = \{No, Yes\}$. $MA = No$ makes reference to a forecasting methodology in which the common factors are fitted with AutoRegressive-Integrated (ARI) models. $MA = Yes$ instead, allows greater complexity since in this case the common factors are modeled as having an AutoRegressive-Integrated-Moving-Average (ARIMA) behavior.
- *Forecast Combinations*: this factor has six levels, $Combinations = \{1, 2, 3, 4, 5, 6\}$ which were described in Section 3.1.

To compare these features, we have performed a computational experiment in which we computed one-day-ahead to two-month-ahead forecasts for every hour and day. This involves estimations for every day during five years (Iberian data), or three and a half years (Italian data), long periods of time that allow validating the results. There are $2 \times 2 \times 2 \times 6 = 48$ treatments resulting from combining all the levels of the aforementioned factors.

We analyze separately the performance of out-of-sample forecasts for various forecasting horizons: one-day-ahead (forecasting horizon $h = 1$), one-week-ahead ($h = 7$), one-month-ahead ($h = 30$), and two-month-ahead ($h = 60$). The goal is to select the treatment which results in the smallest forecasting error possible (measured by MAE) for each of these forecasting horizons h.

Furthermore, and given the fact that forecasts have been computed for a large number of days, the particular *Day* could also explain some significant part of the variability of the response variable, MAE. For instance, if the prices in one day are rather unexpected for being too low or high, the MAE will be large, whatever the values for *Logarithm*, *Historical Length*, *Moving Average (MA)*, and *Forecast Combinations*. Therefore, *Day* is considered as a block in the computational experiment. This helps remove a likely correlation between forecasting errors.

An ANOVA with four factors and one block is conducted to compare the alternative forecasting methodologies (see Montgomery, 1984, for a complete reference on ANOVA and Design of Experiments).

The equation of the model is:

$$MAE_{ijkld} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \epsilon_d + u_{ijkld}, \tag{11}$$

$$u_{ijkld} \sim NIID(0, \sigma_u^2),$$

where μ is the grand mean, and α_i, β_j, γ_k, δ_l, and ϵ_d are known as the main effects of the factors *Logarithm*, *Historical Length*, *Moving Average (MA)*, *Forecast Combinations*, and the block *Day*, respectively. For instance, the main effect δ_l measures the increase or decrease of the average response of the *Forecast Combinations* $l = 1, 2, 3, 4, 5, 6$ with respect to the average level. Similar interpretations apply for the rest of the effects. This is related to the restrictions $\sum_{i=\{No,Yes\}} \alpha_i = 0$, $\sum_{j=\{308,548\}} \beta_j = 0$,

$\sum_{k=\{No,Yes\}} \gamma_k = 0$, $\sum_{l=1}^{6} \delta_l = 0$, and $\sum_{d=1}^{D} \epsilon_d = 0$ (D represents the total number of days with forecasts, this is $D = 1767$ for the Iberian data, and $D = 1219$ for the Italian data).

The noise term u_{ijkld} includes all that is not explicitly taken into account in the model, but that somehow is able to explain some of the variability of the response variable MAE_{ijkld}.

Since it is assumed that the error term u_{ijkld} is Gaussian, independently and identically distributed, with zero mean and variance σ_u^2, once the model has been estimated, a diagnostic checking must be performed, testing that the $\widehat{u}_{ijkld}$ are homoskedastic, Gaussian and independent, where

$$\widehat{u}_{ijkld} = MAE_{ijkld} - \widehat{\mu} - (\widehat{\alpha})_i - (\widehat{\beta})_j - (\widehat{\gamma})_k - (\widehat{\delta})_l - (\widehat{\epsilon})_d.$$

The ANOVA is conducted for each forecasting horizon, and the results are summarized in the next Section 4.2.1, and fully described in Appendices A and B. In all the cases, the response variable was transformed after a first attempt to fit a model to the MAE_{ijkld} because the residuals were heteroskedastic. The results shown hereafter consider the $\ln(MAE_{ijkld})$ as the response variable. Given that the logarithm is a monotonically increasing function, the results can be interpreted directly, and the best model corresponds to the smallest $\ln(MAE_{ijkld})$, while the worst model to the largest $\ln(MAE_{ijkld})$.

Often, in practice the Gaussianity assumption for the ANOVA residuals does not hold. Therefore, the p-values to assess whether the Design of Experiment factor effects are significant or not are recalculated employing bootstrap, following Davison and Hinkley (1997). Likewise, the confidence intervals for the mean of the main effects are obtained employing bootstrap. See Appendix C for further detail on the bootstrap procedures employed.

4.2.1. Summarizing the Conclusions from the ANOVA

For the Iberian data-set, for all the forecasting horizons considered, taking *Logarithm* of prices does not make a difference in performance. Regarding the *Historical Length* of the data, the short window of 308 days is preferred for the forecasting horizons of 1 and 7-day-ahead (forecasts for the short term) while the long window of 548 days is preferred for the forecasting horizons of 30 and 60-day-ahead (long term forecasts); this is consistent with the results in Alonso et al. (2011). Furthermore, the *Moving Average* terms for the factor models are statistically significant for all forecasting horizons, which means that modeling the common factors as ARIMA reduces the error in comparison to modeling them as ARI.

Regarding the *Forecast Combinations*, the median-based combination, mean-based combination and mean BIC-based combination result in better forecasts than the benchmark BIC-selected model and the other combinations available for most forecasting horizons ($h = 7$ onward). However, it is not clear that one of these three is best: the confidence intervals for the median-based combination, mean-based combination and mean BIC-based combination usually overlap, indicating no significant difference between them.

In the case of the Italian data-set, employing *Logarithm* = *Yes* contributes to reduce the forecasting error for all the horizons considered. The *Historical Length* behaves differently than the way it does for the Iberian data-set for $h = 30$, for which case it is convenient to set it to 308 days instead of the 548 days that are suggested for the peninsula. *MA* is also a factor which contributes to reduce the forecasting error, for any forecasting horizon considered. As it occurs with the results for the Iberian electricity prices, *Forecast Combinations* 2, 3 and 6 reveal better results than the benchmark, and it is also not clear that any of the three would be better than the other two in all scenarios. On the contrary, *Combinations* 4 and 5 fail to outperform the benchmark, specially for the long-term forecasting horizons.

For details of the ANOVA results for each forecasting horizon see Appendices A and B.

4.3. Electricity Price Forecasting

In this section, the results of the Forecast Combinations are presented, in comparison with the best model selected by the BIC information criterion[4]. Forecasts are calculated for a long period, for each day and hour. The next paragraphs describe technical details involved in estimation. Subsection 4.3.1 sheds light on the results involved in the estimation of each rolling window and Subsection 4.3.2 presents the results for all days and hours for up to 60-day-ahead forecasts.

Prices are transformed using logarithms to mitigate the existing heteroskedasticity, present in most commodity prices' time series. Therefore, the series modeled are $y_t = ln(P_t + k)$, where P_t represents the vector of 24 prices for day t, and $k = 1,000$.

For medium- and long-term forecasting, forecasts of specific components are negligible. Therefore, we do not model these, but only the unobserved common factors, which explain the larger portion of the variability, and capture the trend of the series in the long-run. This is in line with the results in Alonso et al. (2011). The prediction horizon will vary from 1 to 60 days, and once the factor(s) are modeled and predicted, the loading matrix is used to obtain the forecasts of the original 24-dimensional vector of prices. Then, the out-of-sample performance of the forecasts is evaluated.

We work with rolling windows of *Historical Length*= 548 days, the best length for medium- and long-term forecasts according to the previous section; and estimate one and two common factors for the Iberian data-set, and also a third factor for the Italian data-set.

In each window, 36 alternative seasonal ARIMA$(p,d,q) \times (P,D,Q)s$ models are estimated for each factor: $p = \{1,2,3\}$, $d = \{0\}$, $q = \{1,2,3\}$, $P = \{0,1\}$, $D = \{1\}$, $Q = \{0,1\}$. Weekly seasonality is included in the model, $s = 7$, but yearly seasonality is not. This follows Alonso et al. (2011), who found no improvement in the prediction error when modeling yearly seasonality in the Iberian market, using a similar length of time for the estimation.

Therefore, in the Iberian case there are 36 models that use only one factor and 1,296 models that use two factors; a total of 1,332 different models, depending on how many factors they include and the parameters of the ARIMA$(p,d,q) \times (P,D,Q)s$. For the Italian data-set, because up to three common factors are estimated, the total number of models for any rolling window is 47,988 (36 models that use only one factor, 1,296 models that use two factors, and 46,656 models that use three factors). These figures make it unfeasible to check the residuals' behavior for each ARIMA model estimated; notwithstanding, TRAMO, the software employed to calculate the ARIMA models, estimates the p-value of the Ljung Box statistic for each model, and shows acceptable values for most cases. Additionally, three of the five Forecast Combinations under consideration are based on BIC, so "badly" behaved models (poor fit will be associated to a high residuals variance) will be assigned small or negligible weights in the final forecast. Furthermore, the median-based combination is not affected by outliers due to "badly" behaved models. Only the mean combination may be affected by them but, based on Tables 3 and 4, median and mean combinations reveal similar results. If there were fewer models or the analyses were limited to a shorter period, residual checking could be performed before forecast averaging. In that case, it would be reasonable to obtain slightly better results.

Notwithstanding the large number of models, the estimation for each individual window of time takes only a few minutes; therefore, the procedure could be used in real time. Additionally, even though with such a large number of models some will be superfluous, the combinations that use weights depending on the BIC will assign them nearly null weights.

For the Iberian market, the complete data-set comprises the period July, 2006, to December, 2012. The data before 2008 is only used as historical data, therefore the first predicted day is January 1st, 2008, and the last one December 31st, 2012. Thus, there is a total of 1,767 time rolling windows,

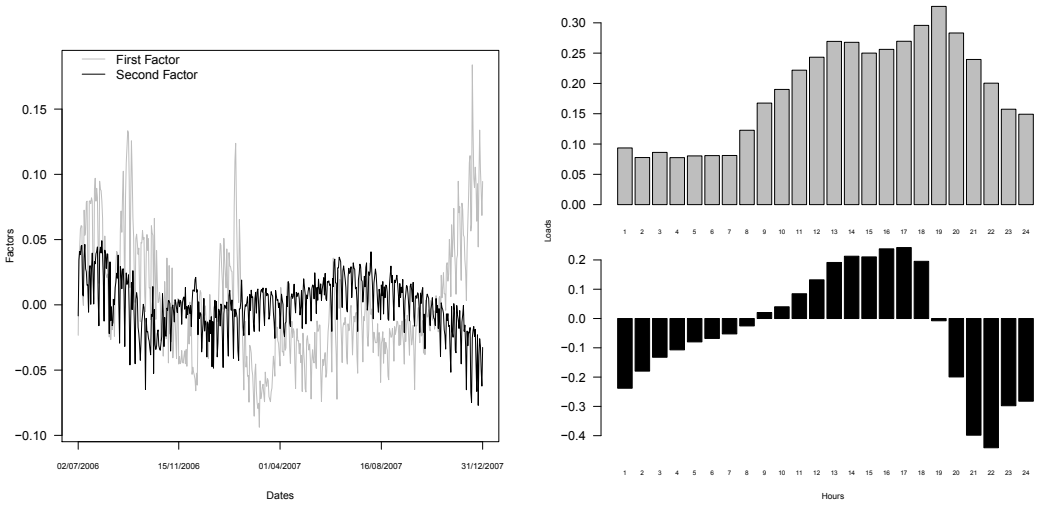
[4] This model may have one or two common factors for the Iberian data, and a third factor as well for the Italian case, and each common factor extracted is modeled with a seasonal ARIMA model with parameters selected by BIC. The model is selected anew in each window.

corresponding to 1,827 days in January 1st, 2008-December 31st, 2012 minus 60 days needed for out-of-sample data (used to compare with up to two-month-ahead forecasts). This data is provided by the market operator, OMIE (Iberian Market Operator). For the Italian market, on the other hand, the complete data-set includes January, 2008 to December, 2012. Therefore, the first predicted day is July 2nd, 2009, and the last one December 31st, 2012. There are 1,219 rolling windows in total corresponding to 1,279 days in July 2nd, 2008-December 31st, 2012 minus 60 days needed for out-of-sample data. The prices for the Italian electricity market are available in the website of the market operator, GME (The Energy Markets Operator).

4.3.1. Illustration for a Single Forecasting Window

Before proceeding with the presentation of the results, this subsection is used to gain insight into the role of the common factors, as well as the forecasting combinations. With this aim, the estimation for one window of the Iberian data-set is analyzed in further detail.

The role of the underlying factors is hereby clarified. Considering as an example the first rolling window in the estimation for the Iberian data-set, Figure 2 presents the first and second common factors, as well as the weights assigned to them for each of the 24 hours. For the first factor, which explains 64.6% of the data variability, weights are heavy from hours 8 to 24, when most people are awake. Then, it is possible to interpret that this factor mainly records the general behavior of prices during hours when people are awake. On the other hand, the weights of the second factor are positive from 9 to 18 and negative otherwise. This coincides with usual working hours or, alternatively, sunlight hours. Therefore, the second factor, which accounts for 13.8% of the data variability, would capture changes in the price relation of working vs. non working hours. Notice that some models would include only the first factor, while others will include the first and second common factors. Models with more factors have been excluded from the analysis for the Iberian electricity prices (setting $r \leq 2$) because already around 80% of the data variability is explained by two factors.



(a) First and second estimated common factors. **(b)** Estimated loads for first (top) and second (bottom) common factors.

Figure 2. Common factors and their weighs corresponding to the first rolling window of the Iberian data-set, with *Historical Length*= 548 days.

The massive estimations performed make it unfeasible to provide with the estimation results for each of the 1,332 models and for each of the 1,767 rolling windows of time. However, as an illustration, for the first rolling window, taking for example the first factor and the ARIMA model of

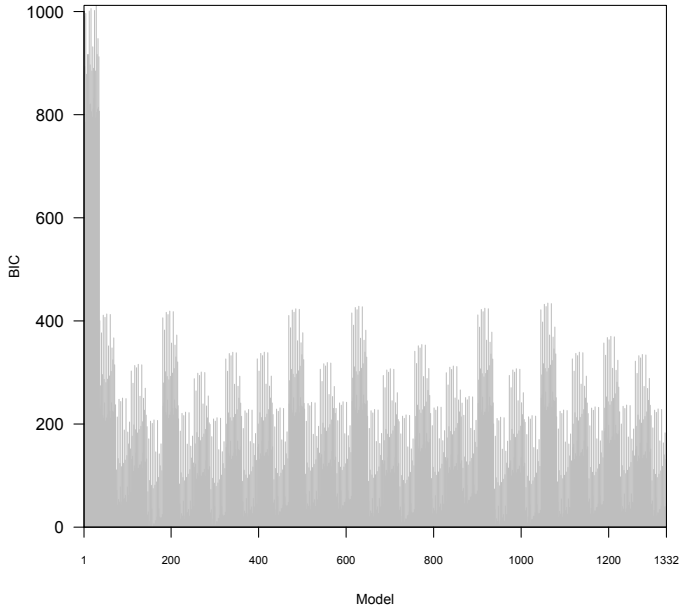


Figure 3. BIC criteria corresponding to models with one and two common factors of the first rolling window of the Iberian data-set, with *Historical Length* = 548 days.

order $p = 1$, $q = 1$, $P = 0$, $Q = 0$, the coefficients are the following: $\phi = 0.7145$, $\theta = -0.1319$, both significant.

To shed some light on how the alternative models enter the combinations, Figure 3 presents the BIC values for the 1,332 previously mentioned models. For illustrative purposes, also the first rolling window of the Iberian data is employed. The horizontal axis corresponds to the indexes of the models. The first 36 values (*X axis* from 1 to 36) represent models with only one factor ($r = 1$), starting with parameters $p = 1$, $q = 1$, $P = 0$, $Q = 0$ for $X\ axis = 1$, then $p = 1$, $q = 1$, $P = 0$, $Q = 1$ for $X\ axis = 2$, until $p = 3$, $q = 3$, $P = 1$, $Q = 1$ for $X\ axis = 36$. *X axis* 37 to 1,336 correspond to models with two factors ($r = 2$), we can see an important reduction of the BIC for these models. In $X\ axis = 1,336$ the order of the ARIMA models for the two factors coincide, $p = 3$, $q = 3$, $P = 1$, $Q = 1$. For BIC dependent combinations, the smaller the BIC value, the greater that model's weight. In this way, better performing models are rewarded. It is clear that there are some models with predominant low BIC (i.e. high weights). Of course, if all considered models had poor goodness of fit, then it would be reasonable that the combinations would inherit that bad performance. The claim in this work is that those combinations would be, at least, as good as the best considered model.

4.3.2. Forecasting Results

Given that, according to the ANOVA, many of the combinations resulted significantly better than the benchmark, in this section we study more closely those improvements. As it was previously indicated, in this section we work with DOE's factors that have the following characteristics: *Logarithm* = *Yes*, *Historic Length* = 548 days, and *MA* = *Yes*. We will consider all the *Combinations*.

Tables 1 and 2 present some descriptive statistics of interest for the daily average MAE (expression (8)). As expected, the error increases with the forecasting horizon h, for all the forecasts available, and for both markets.

For $h = 1$, the *Combinations* do not usually do better than the benchmark model, either comparing means or quartiles. Though this may seem contradictory to the ANOVA's findings, it is not: as we explained in subsection 4.3.1, we are not employing the suggested values for the DOE factors for short-run forecasts. The ANOVA's outcomes indicate an advantage of using the short *Historic Length*, which we do not do here. The reason for this is to focus on the performance of a particular specification, which in this case reflects an interest in medium- and long-run forecasts rather than short-run forecasts.

On the contrary, we find that, for longer forecasting horizons ($h \geq 7$), the *Combinations* consistently render smaller errors than the benchmark. In particular, *Combinations* 2, 3 and 6 perform well in these horizons, for both data-sets.

Table 1. Descriptive statistics for MAE. Iberian Market.

Forecasting horizon		BIC-selected model	Median-based Combination	Mean-based Combination	BIC-based Combination	BIC 50% Combination	Mean-BIC-based Combination
h=1	mean	5.3389	5.3617	5.4107	5.3346	5.3346	5.3276
	Q1	3.2591	3.2709	3.2850	3.2632	3.2632	3.2496
	Q2	4.5014	4.5198	4.5973	4.4970	4.4970	4.4989
	Q3	6.3009	6.3391	6.3991	6.2910	6.2910	6.2734
h=7	mean	6.3261	6.1562	6.1648	6.3142	6.3142	6.1400
	Q1	3.6423	3.5485	3.5326	3.6457	3.6457	3.5344
	Q2	5.1323	4.9770	4.9612	5.1261	5.1261	4.9754
	Q3	7.4496	7.4044	7.3966	7.4500	7.4500	7.3105
h=30	mean	7.8677	7.5395	7.5531	7.8411	7.8411	7.4725
	Q1	4.4447	4.2406	4.2163	4.4416	4.4416	4.1877
	Q2	6.3851	6.1435	6.1057	6.3668	6.3668	6.0195
	Q3	9.8775	9.5081	9.5525	9.8372	9.8372	9.3183
h=60	mean	9.5120	9.1545	9.1496	9.4938	9.4938	9.0772
	Q1	5.2904	4.8470	4.8116	5.2651	5.2651	4.7933
	Q2	7.7493	7.2159	7.2676	7.7290	7.7290	7.2704
	Q3	11.5845	11.4665	11.4378	11.5386	11.5386	11.2074

For evaluating the performance of the *Combinations* in direct comparison to the benchmark we use the relative MAE (RelMAE). This is presented in Figures 4a and 4b. For most forecast horizons considered, all combinations of forecasts included hereby outperform the best factor model selected by the BIC ($RelMAE < 1$). Though for the very short time, the median-based and mean-based *Combinations* are worse than the benchmark, they outperform it in the medium- and long-run, which is the aim of this section's exercise. The mean BIC-based Combination is better than the others for all forecasting horizons, and for both data-sets. After a certain h, the performance of the *Combinations* relatively to the benchmark becomes stable as the forecasting horizon increases, an advantage when the focus is obtaining accurate forecasts in the long-run. On the contrary, the vast majority of methods proposed in the literature are only appropriate for short-term forecasting, because their performance dramatically degrades when extending the forecasting horizon.

Table 2. Descriptive statistics for MAE. Italian Market.

Forecasting horizon		BIC-selected model	Median-based Combination	Mean-based Combination	BIC-based Combination	BIC 50% Combination	Mean-BIC-based Combination
h=1	mean	7.7263	7.8204	7.8829	7.7198	7.7198	7.7104
	Q1	4.9388	4.9486	4.9366	4.9253	4.9253	4.8800
	Q2	6.5024	6.5837	6.5857	6.5053	6.5053	6.4864
	Q3	9.0010	9.2146	9.3917	9.0167	9.0167	9.0471
h=7	mean	8.9553	8.8204	8.8323	8.9451	8.9451	8.7682
	Q1	5.3898	5.2714	5.2710	5.3849	5.3849	5.3467
	Q2	7.3988	7.2220	7.2518	7.4005	7.4005	7.2698
	Q3	10.4787	10.3937	10.3449	10.5152	10.5152	10.1510
h=30	mean	10.6954	10.4080	10.4516	10.6815	10.6815	10.3378
	Q1	6.3428	6.1296	6.1021	6.3335	6.3335	6.1352
	Q2	8.9068	8.4948	8.5843	8.9069	8.9069	8.4545
	Q3	13.1087	12.5893	12.7182	13.0952	13.0952	12.4662
h=60	mean	12.2726	11.7872	11.8268	12.2225	12.2225	11.7028
	Q1	7.3961	7.2301	7.3933	7.4088	7.4088	7.2269
	Q2	10.6249	10.0720	10.0532	10.5705	10.5705	10.0262
	Q3	15.1359	14.2388	14.1938	14.9273	14.9273	14.1697

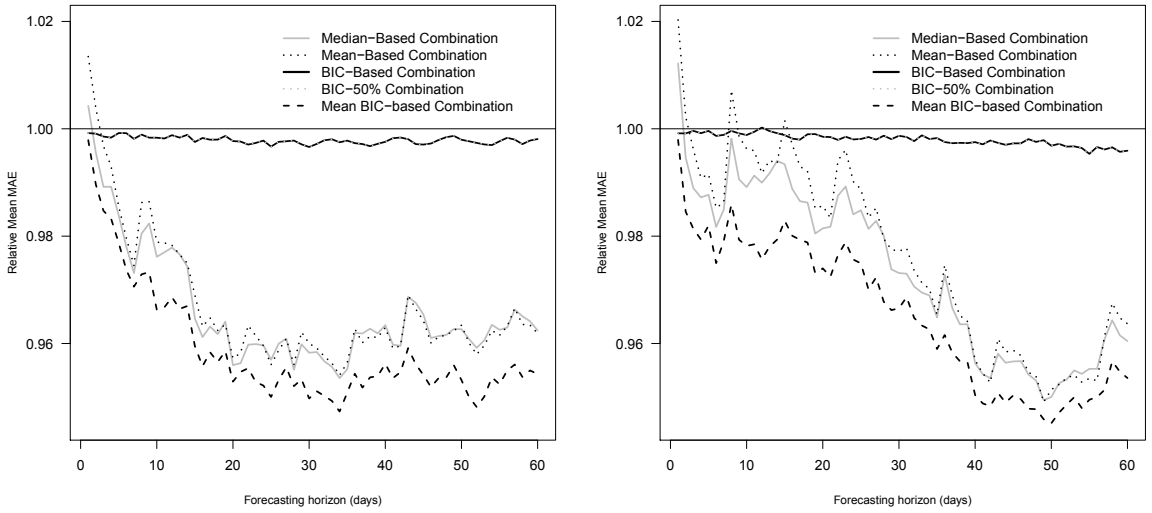
Table 3. Weekly, monthly, and bi-monthly MAE and MedAE for the Iberian Market.

	BIC-selected model	Median-based Combination	Mean-based Combination	BIC-based Combination	BIC 50% Combination	Mean-BIC-based Combination
Weekly						
MAE	5.9455	5.8690	5.8965	5.9384	5.9384	5.8397
MedAE	5.3515	5.2433	5.2677	5.3444	5.3444	5.2275
Monthly						
MAE	6.9069	6.6952	6.7097	6.8934	6.8934	6.6526
MedAE	6.3635	6.1179	6.1367	6.3484	6.3484	6.0882
Bi-Monthly						
MAE	7.8184	7.5456	7.5512	7.8014	7.8014	7.4867
MedAE	7.3047	7.0081	7.0157	7.2844	7.2844	6.9539

Last, in Tables 3 and 4, the MAE and MedAE for the BIC-selected model and for the alternative *Combinations* are presented for weekly, monthly and bi-monthly forecasts[5]. Results are consistent with those of the previous tables. Similar outcomes were obtained with a different accuracy metric, the Root Mean Squared Error (RMSE, see Appendix D for details) (Hyndman and Koehler, 2006).

In conclusion, there is an improvement in prediction when using *Forecast Combinations*, specially median-based, mean-based, and median-BIC-based *Combinations*, in comparison with the best model selected according to the BIC criterion. Even though the decrease in the errors is small, it is steady, supporting the conclusion obtained in the ANOVA, in which some combinations are statistically significant better than the benchmark.

[5] Weekly values are obtained by dividing the average of the week (the MAE of horizons $h = 1$ to $h = 7$) by the forecasting horizon ($h = 7$). Similarly for the monthly and bi-monthly values.



(a) Iberian market.

(b) Italian market.

Figure 4. Relative MAE, forecast horizon from 1- to 60-day-ahead. Values smaller than 1 indicate a result outperforming the benchmark.

Table 4. Weekly, monthly, and bi-monthly MAE and MedAE for the Italian Market.

	BIC-selected model	Median-based Combination	Mean-based Combination	BIC-based Combination	BIC 50% Combination	Mean-BIC-based Combination
Weekly						
MAE	8.5488	8.4687	8.5109	8.5418	8.5418	8.3989
MedAE	7.2800	7.2110	7.2525	7.2752	7.2752	7.1404
Monthly						
MAE	9.7019	9.5744	9.6197	9.6902	9.6902	9.4794
MedAE	8.3685	8.2531	8.3021	8.3579	8.3579	8.1718
Bi-Monthly						
MAE	10.6000	10.3025	10.3310	10.5779	10.5779	10.2186
MedAE	9.1880	8.9105	8.9379	9.1656	9.1656	8.8332

5. Conclusions and further lines of research

In this paper, Dynamic Factor Models and Forecast Combination techniques have been jointly employed to obtain predictions of spot market electricity prices in the Iberian and Italian Markets. The main contribution consists, therefore, of combining two streams of literature in order to obtain forecasts that outperform those resulting from the individual models. In this respect, there are three combinations that clearly outperform the benchmark: the median-based combination, the mean-based combination, and the mean BIC-based combination. This conclusion is supported by the ANOVA of the combinations for forecast horizons 1-day-ahead, 7-day-ahead, 30-day-ahead and 60-day-ahead.

In the process of trying to obtain the best possible results, different aspects of the available models were compared. In this regard, the main conclusions are that longer historic data-sets benefit longer forecasting horizons, and the error is reduced by the inclusion of MA terms when modeling the unobserved factors (vs. AR models).

This application reflects how the methodology works in empirical applications. The numerical results for electricity prices, which is a difficult to predict series, are good. An effort has been made to obtain the results for many time horizons ($h = 1$ to $h = 60$), for every day and during several years, and considering several models for the factors, enhancing the validity of our proposal. In order words, forecasts are obtained for the very short (one day) and short term (a few days ahead), like most of other works, as well as for the medium term, which is an extension not customary in the literature. As previously explained, this approach can be employed to obtain long term forecasts (even up to a year) not experiencing a degradation of accuracy, which is a drawback that most applications suffer from.

Numerous lines of research remain open in relation with this topic. For instance, in this work, few techniques for combining forecasts are employed besides the mean, and weights depend on the overall performance of the particular model to be used in the combination in terms of the BIC information criterion. However, there are several other, Bayesian and classical techniques to determine such weights. In particular, it would be interesting to compare the performance of both types of techniques. Moreover, in this article we have worked with fixed weights; however, these could change in a predefined way for different forecasting horizons. Furthermore, weights could be adaptive to the performance of the models (as in Sánchez, 2006).

The use of ARIMA models for the common factors allows to maintain the number of parameters to estimate low, but it may also signify a constraint in the improvement that can be achieved from the combinations of forecasts. A future line of work would be to include other models for the factors, such as the SeaDFA, which assumes that vector F_t follows a VARIMA model, or the Generalized AutoRegressive Conditionally Heteroscedastic (GARCH)-SeaDFA.

It is also left for future work to incorporate in the Forecast Combination other forecasting methods, not necessarily involving DFM. For example, the predictions obtained by García-Martos et al. (2007) mixed model, which presents extremely accurate short-term predictions for the Iberian market. With weights evolving for different time-horizons, including this model for short-term predictions could improve the results.

A further improvement could consist of employing explanatory variables that drive spot prices in the models. Some examples of these variables are demand, weather conditions, fuel prices, production by technology, and excess capacity. Interesting references are Bello et al. (2016), Monteiro et al. (2015), and Di Cosmo (2015). However, it would be necessary to assess if the uncertainty in the prediction of the explanatory variables does not outweigh the improvement in the forecast of the price. In a similar line of research, regime switching models could be employed to deal with spikes in the price series.

Last, bootstrap procedures could be used to obtain confidence intervals of the predictions and in this way assess the uncertainty involved in the forecasts.

Table A1. Summary of results of the Analysis of variance for ln(MAE) for all forecasting horizons. Iberian market.

	$h=1$	$h=7$	$h=30$	$h=60$
Logarithm	ns	ns	ns	ns
Hist Length (days)	308	308	548	548
MA	Yes	Yes	Yes	Yes
Combinations	$\{6\}$	$\{2,3,6\}$	$\{2,3,6\}$	$\{2,3,6\}$

Notes: Significance level of at least $\alpha = 0.01$. ns: not significant.

Table A2. Analysis of variance for ln(MAE). Main effects. Forecast horizon $h = 1$, Iberian market.

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Logarithm	1	0.02	0.02	0.67	0.4331
Hist Length	1	1.56	1.56	59.45	0.0020
MA	1	111.24	111.24	4240.25	0.0020
Combinations	5	0.35	0.07	2.70	0.0140
Day	1766	22166.97	12.55	478.46	0.0020
Residuals	83041	2178.51	0.03		

Notes: DF stands for Degrees of Freedom. All F-ratios are based on the residual mean square error. P-values are estimated by bootstrap.

Acknowledgments:

A.M. Alonso acknowledges support of the Ministry of Economy and Competitiveness, Spain, by projects ECO2012-38442, and ECO2015-66593.

Carolina García-Martos acknowledges financial support from project DPI2011-23500, Ministry of Economy and Competitiveness, Spain.

The authors would like to extend their appreciation to Professor Michael Wiper for his assistance and corrections regarding the proper use of English in this document.

Appendix A

Appendix A.1. Details of ANOVA for a Comparison of the Alternatives for Modeling. Results for the Iberian Electricity Market.

In this section, we describe in detail the results for the ANOVA performed for the forecasting horizons $h = 1, 7, 30, 60$, for the data-set of prices in the Iberian electricity market. In order to be robust against departures from the Gaussianity assumption, we employ bootstrap to calculate the ANOVA's p-values, as well as the confidence intervals for the means of the DOE's factors.

Table A1 presents a summary of the results described in the following sections. For each DOE's factor it indicates the best level. For instance, for $h = 1$, *Historical Length*= 308 days outperforms the alternative *Historical Length*= 548 days. We work with $ln(MAE)$ as dependent variable in order to eliminate heteroskedasticity.

Appendix A.1.1. Minimizing Forecasting Error for One-Day-Ahead Forecasts ($h = 1$)

To assess the results for one-day-ahead forecasts ($h = 1$), see Figure A1 and Table A2. In Figure A1, the horizontal axis presents the alternative values of the DOE's factors (*Logarithm*, *Historical Length*, *Moving Average*, *Forecast Combinations*), and the vertical axis shows the logarithm of MAE corresponding to the means and 95% confidence intervals.

The effect of DOE factor *Logarithm* is not significant. On the contrary, when considering *Historical Length*, we can see a significant difference: using the short historic window gives significantly better forecasts in terms of forecasting accuracy - a smaller MAE - than using the long historic window (548 days). Regarding the *Moving Average* component, significantly better results are obtained when incorporating an MA term in the model of the unobserved common factors (the alternative being modeling as ARI). Last, *Forecast Combination* 6 (mean BIC-based combination) outperforms most of the others.

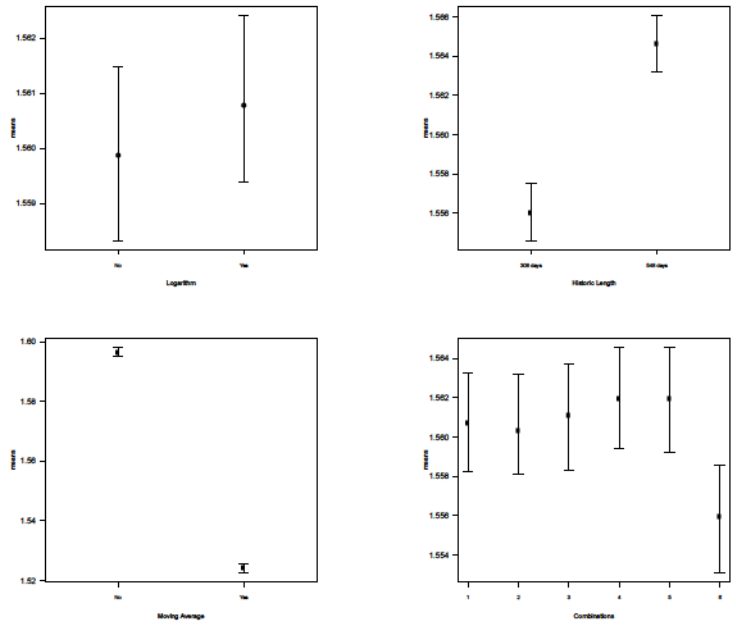


Figure A1. Bootstrap confidence intervals for the mean $ln(MAE)$ of the factors *Logarithm*, *Historical Length*, *Moving Average*, and *Forecast Combinations*. *Forecast Combinations* include: (1) benchmark BIC-selected model, (2) median-based combination, (3) mean-based combination, (4) BIC-based combination, (5) BIC-50% combination, (6) mean BIC-based combination. Forecast horizon $h = 1$, Iberian market.

Table A3. Analysis of variance for ln(MAE). Main effects. Forecast horizon $h = 7$, Iberian market.

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Logarithm	1	0.07	0.07	2.66	0.1218
Hist Length	1	0.60	0.60	23.07	0.0020
MA	1	7.71	7.71	295.44	0.0020
Combinations	5	8.37	1.67	64.18	0.0020
Day	1766	25363.03	14.36	550.59	0.0020
Residuals	83041	2166.09	0.03		

Notes: DF stands for Degrees of Freedom. All F-ratios are based on the residual mean square error. P-values are estimated by bootstrap.

Appendix A.1.2. Minimizing Forecasting Error for Seven-Day-Ahead Forecasts ($h = 7$)

Table A3 and Figure A2 present the results for seven-day-ahead forecasts ($h = 7$). There are three *Forecast Combinations* which outperform the benchmark: 2, 3 and 6. Furthermore, there is a significant difference between the two values for *Historical Length*: employing historic data-sets of 308 days provides with significantly better forecasts than 548 days. Significantly better results are obtained when incorporating a *Moving Average* component in the unobserved common factors' models. Last, the effect of *Logarithm* continues to be not significant.

Appendix A.1.3. Minimizing Forecasting Error for One-Month-Ahead Forecasts ($h = 30$)

Details on the results for thirty-day-ahead forecasts ($h = 30$) can be found in Table A4 and Figure A3. For *Forecast Combinations*, we obtain similar results to $h = 7$. Contrary to shorter forecasting horizons, the *Historical Length* of 548 days presents significantly better forecasts than the shorter window, an intuitive result. Again, significantly better results are obtained with a *Moving Average* component in the model for unobserved common factors. The effect of *Logarithm* continues to be not significant.

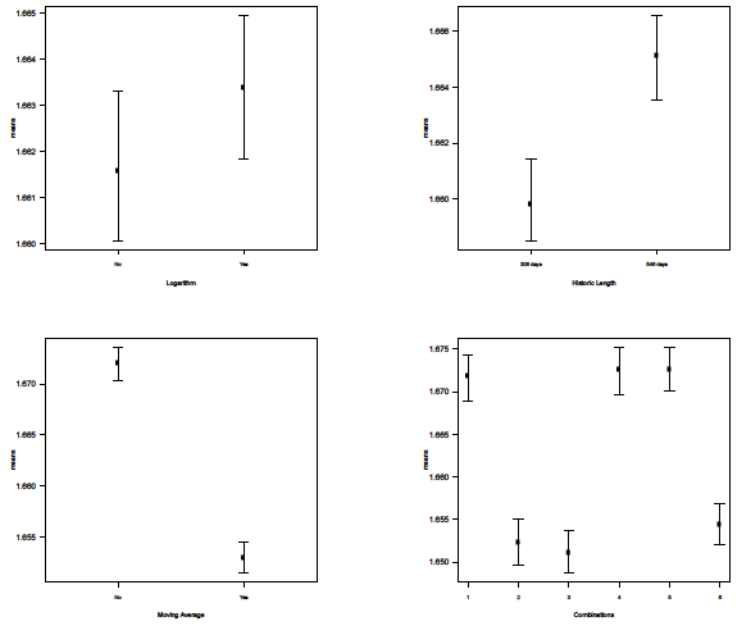


Figure A2. Bootstrap confidence intervals for the mean $ln(MAE)$ of the factors *Logarithm*, *Historical Length*, *Moving Average*, and *Forecast Combinations*. *Forecast Combinations* include: (1) benchmark BIC-selected model, (2) median-based combination, (3) mean-based combination, (4) BIC-based combination, (5) BIC-50% combination, (6) mean BIC-based combination. Forecast horizon $h = 7$, Iberian market.

Table A4. Analysis of variance for ln(MAE). Main effects. Forecast horizon $h = 30$, Iberian market.

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Logarithm	1	0.01	0.01	0.33	0.5669
Hist Length	1	19.90	19.90	458.58	0.0020
MA	1	10.34	10.34	238.25	0.0020
Combinations	5	11.35	2.27	52.31	0.0020
Day	1766	24319.88	13.77	317.31	0.0020
Residuals	83041	3604.00	0.04		

Notes: DF stands for Degrees of Freedom. All F-ratios are based on the residual mean square error. P-values are estimated by bootstrap.

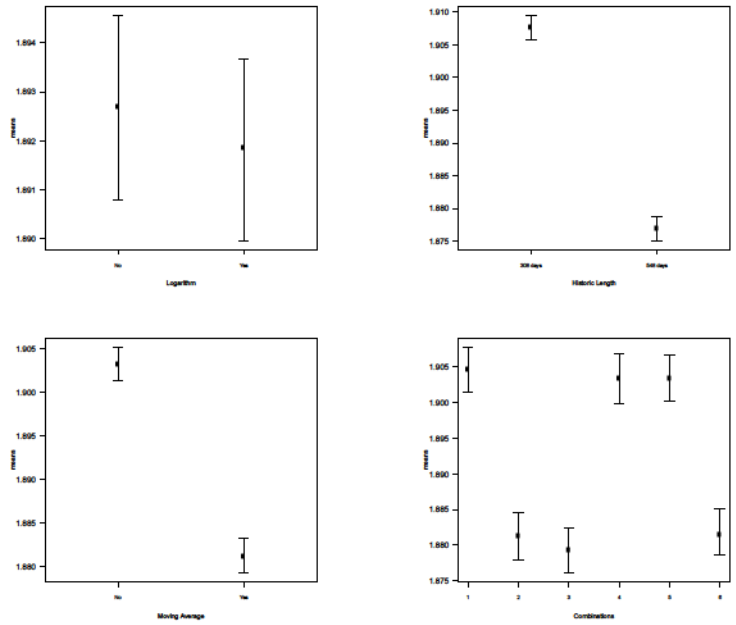


Figure A3. Bootstrap confidence intervals for the mean $ln(MAE)$ of the factors *Logarithm*, *Historical Length*, *Moving Average*, and *Forecast Combinations*. *Forecast Combinations* include: (1) benchmark BIC-selected model, (2) median-based combination, (3) mean-based combination, (4) BIC-based combination, (5) BIC-50% combination, (6) mean BIC-based combination. Forecast horizon $h = 30$, Iberian market.

Table A5. Analysis of variance for ln(MAE). Main effects. Forecast horizon $h = 60$, Iberian market.

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Logarithm	1	0.01	0.01	0.15	0.6946
Hist Length	1	126.98	126.98	2359.20	0.0020
MA	1	6.73	6.73	125.11	0.0020
Combinations	5	20.82	4.16	77.37	0.0020
Day	1766	25660.02	14.53	269.97	0.0020
Residuals	83041	4469.42	0.05		

Notes: DF stands for Degrees of Freedom. All F-ratios are based on the residual mean square error. P-values are estimated by bootstrap.

Appendix A.1.4. Minimizing Forecasting Error for Two-Month-Ahead Forecasts ($h = 60$)

The longest forecasting horizon considered in this assessment is sixty-day-ahead ($h = 60$). The results are presented in Table A5 and Figure A4. As for $h = 7$ and $h = 30$, we obtain that the three most successful *Forecast Combinations* are 2, 3 and 6. Employing a *Historical Length* of 548 days provides with significantly better forecasts in terms of forecasting accuracy than using the shorter option. Additionally, significantly better results are obtained when incorporating a *Moving Average* component to model the unobserved common factors. Last, there is no change with respect to the conclusions for *Logarithm*.

Appendix B.

Appendix B.1. Details of ANOVA for a Comparison of the Alternatives for Modeling. Results for the Italian Electricity Market.

In this section, we describe in detail the results for the ANOVA performed for the forecasting horizons $h = 1, 7, 30, 60$, for the data-set of prices in the Italian electricity market. In order to be

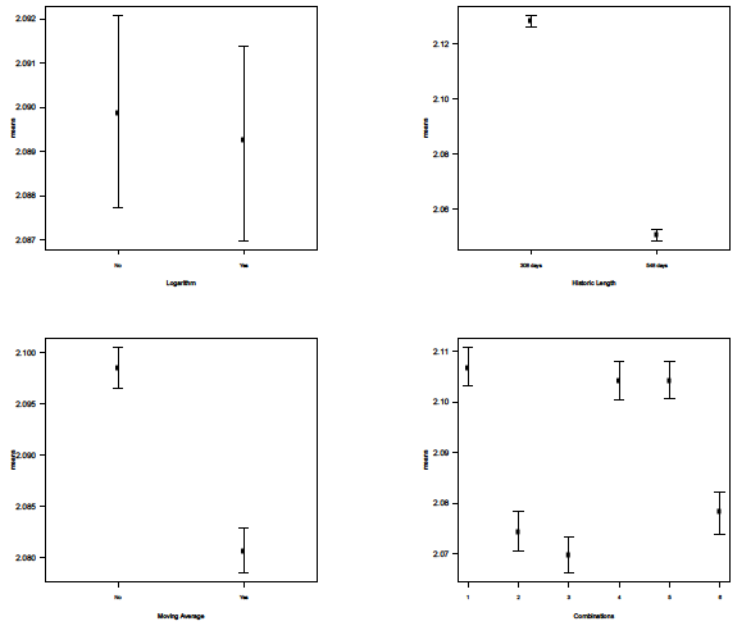


Figure A4. Bootstrap confidence intervals for the mean $ln(MAE)$ of the factors *Logarithm*, *Historical Length*, *Moving Average*, and *Forecast Combinations*. *Forecast Combinations* include: (1) benchmark BIC-selected model, (2) median-based combination, (3) mean-based combination, (4) BIC-based combination, (5) BIC-50% combination, (6) mean BIC-based combination. Forecast horizon $h = 60$, Iberian market.

Table B1. Summary of results of the Analysis of variance for ln(MAE) for all forecasting horizons. Italian market.

	$h = 1$	$h = 7$	$h = 30$	$h = 60$
Logarithm	Yes	Yes	Yes	Yes
Hist Length (days)	308	308	308	548
MA	Yes	Yes	Yes	Yes
Combinations	$\{2:6\}$	$\{2:6\}$	$\{2,3,6\}$	$\{6\}$

Notes: Significance level of at least $\alpha = 0.01$. ns: not significant.

robust against departures from the Gaussianity assumption, we employ bootstrap to calculate the p-values of the analysis, as well as the confidence intervals for the means of the DOE's factors.

Table B1 presents a summary of the results described in the following sections. Notice that we work with $ln(MAE)$ for dependent variable in order to eliminate heteroskedasticity.

Appendix B.1.1. Minimizing Forecasting Error for One-Day-Ahead Forecasts ($h = 1$)

For the shortest forecasting horizon considered ($h = 1$), we obtain that all the factors included in the DOE affect the forecasting error measured by $ln(MAE)$ (see Table B2). Figure B1 presents 95% confidence intervals showing the effect of the DOE factors in the mean of $ln(MAE)$. Better results are obtained when the dependent variable is $ln(Prices)$ rather than *Prices*, and also when the short *Historical Length* is used. Employing ARIMA models for the common factors instead of ARI models also results in a smaller forecasting error. We find that all the *Forecast Combinations* proposed, specially 6, turn out to provide significantly better results than the benchmark (presented as *Combination* 1 in the plot).

Table B2. Analysis of variance for ln(MAE). Main effects. Forecast horizon $h = 1$, Italian market.

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Logarithm	1	49.61	49.61	1092.13	0.0020
Hist Length	1	25.78	25.78	567.39	0.0020
MA	1	220.14	220.14	4846.00	0.0020
Combinations	5	5.66	1.13	24.91	0.0020
Day	1218	12469.53	10.24	225.36	0.0020
Residuals	57285	2602.32	0.05		

Notes: DF stands for Degrees of Freedom. All F-ratios are based on the residual mean square error. P-values are estimated by bootstrap.

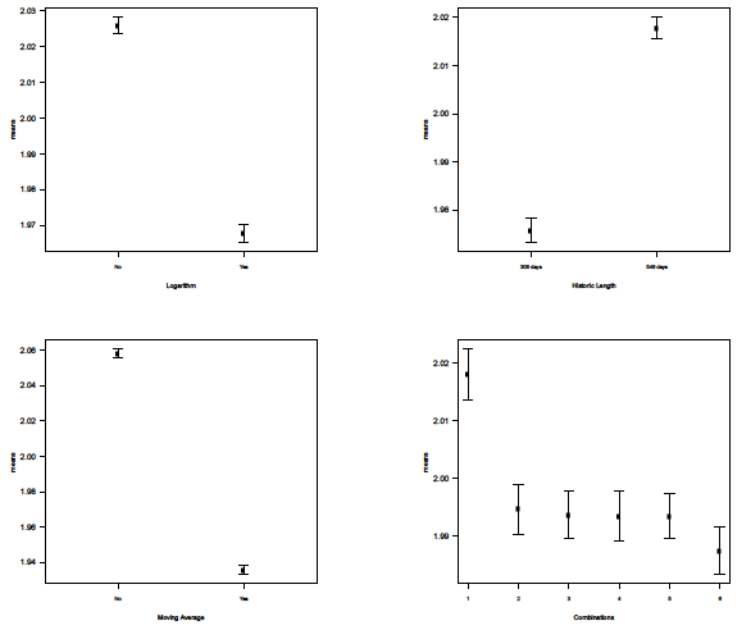


Figure B1. Bootstrap confidence intervals for the mean $ln(MAE)$ of the factors *Logarithm*, *Historical Length*, *Moving Average*, and *Forecast Combinations*. *Forecast Combinations* include: (1) benchmark BIC-selected model, (2) median-based combination, (3) mean-based combination, (4) BIC-based combination, (5) BIC-50% combination, (6) mean BIC-based combination. Forecast horizon $h = 1$, Italian market.

Table B3. Analysis of variance for ln(MAE). Main effects. Forecast horizon $h = 7$, Italian market.

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Logarithm	1	17.17	17.17	530.14	0.0020
Hist Length	1	10.42	10.42	321.78	0.0020
MA	1	60.84	60.84	1878.39	0.0020
Combinations	5	10.03	2.01	61.94	0.0020
Day	1218	13982.28	11.48	354.44	0.0020
Residuals	57285	1855.34	0.03		

Notes: DF stands for Degrees of Freedom. All F-ratios are based on the residual mean square error. P-values are estimated by bootstrap.

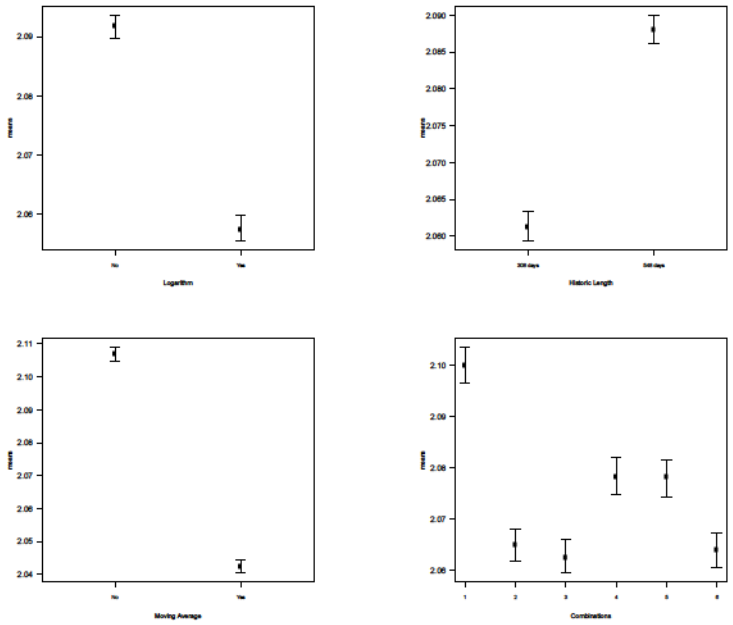


Figure B2. Bootstrap confidence intervals for the mean $ln(MAE)$ of the factors *Logarithm*, *Historical Length*, *Moving Average*, and *Forecast Combinations*. *Forecast Combinations* include: (1) benchmark BIC-selected model, (2) median-based combination, (3) mean-based combination, (4) BIC-based combination, (5) BIC-50% combination, (6) mean BIC-based combination. Forecast horizon $h = 7$, Italian market.

Appendix B.1.2. Minimizing Forecasting Error for Seven-Day-Ahead Forecasts ($h = 7$)

For $h = 7$, all the factors included in the DOE affect the forecasting error measured by $ln(MAE)$ (see Table B3). The confidence intervals showing the effect of the DOE's factors in the mean of $ln(MAE)$ are presented in Figure B2. Similar results to $h = 1$ are obtained for *Logarithm*, *Historical Length*, *Moving Average*, and *Forecast Combinations*. Especially *Forecast Combinations* 2, 3 and 6, turn out to provide significantly better results than the benchmark *Forecast Combination* 1.

Appendix B.1.3. Minimizing Forecasting Error for One-Month-Ahead Forecasts ($h = 30$)

For $h = 30$, we obtain similar results to those presented for shorter horizons (see Table B4). Figure B3 presents the confidence intervals showing the effect of the factors in the mean of $ln(MAE)$. In the case of *Forecast Combinations*, 4 and 5 no longer provide a significant advantage.

Table B4. Analysis of variance for ln(MAE). Main effects. Forecast horizon $h = 30$, Italian market.

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Logarithm	1	21.23	21.23	453.13	0.0020
Hist Length	1	6.21	6.21	132.49	0.0020
MA	1	140.73	140.73	3004.28	0.0020
Combinations	5	2.36	0.47	10.09	0.0020
Day	1218	13557.33	11.13	237.62	0.0020
Residuals	57285	2683.40	0.05		

Notes: DF stands for Degrees of Freedom. All F-ratios are based on the residual mean square error. P-values are estimated by bootstrap.

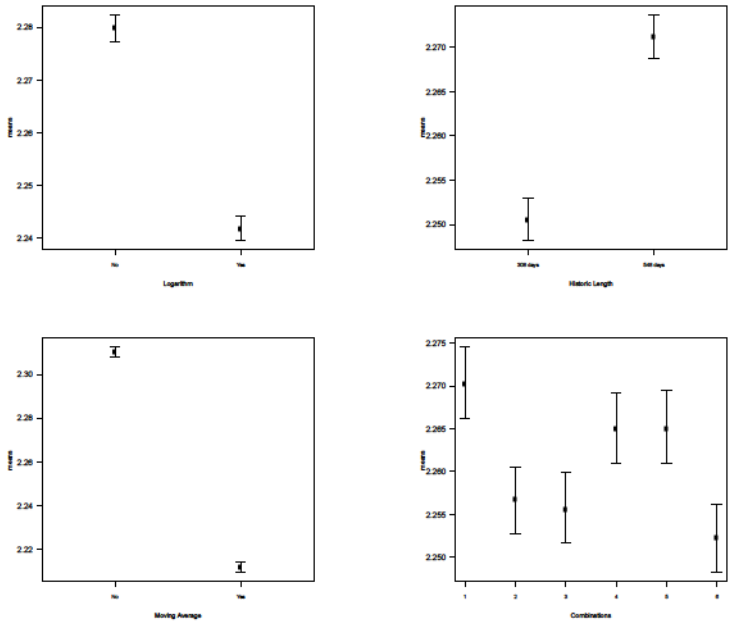


Figure B3. Bootstrap confidence intervals for the mean $ln(MAE)$ of the factors *Logarithm*, *Historical Length*, *Moving Average*, and *Forecast Combinations*. *Forecast Combinations* include: (1) benchmark BIC-selected model, (2) median-based combination, (3) mean-based combination, (4) BIC-based combination, (5) BIC-50% combination, (6) mean BIC-based combination. Forecast horizon $h = 30$, Italian market.

Table B5. Analysis of variance for ln(MAE). Main effects. Forecast horizon $h = 60$, Italian market.

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Logarithm	1	5.21	5.21	110.18	0.0020
Hist Length	1	1.93	1.93	40.86	0.0020
MA	1	83.70	83.70	1770.80	0.0020
Combinations	5	1.69	0.34	7.17	0.0020
Day	1218	13013.62	10.68	226.05	0.0020
Residuals	57285	2707.58	0.05		

Notes: DF stands for Degrees of Freedom. All F-ratios are based on the residual mean square error. P-values are estimated by bootstrap.

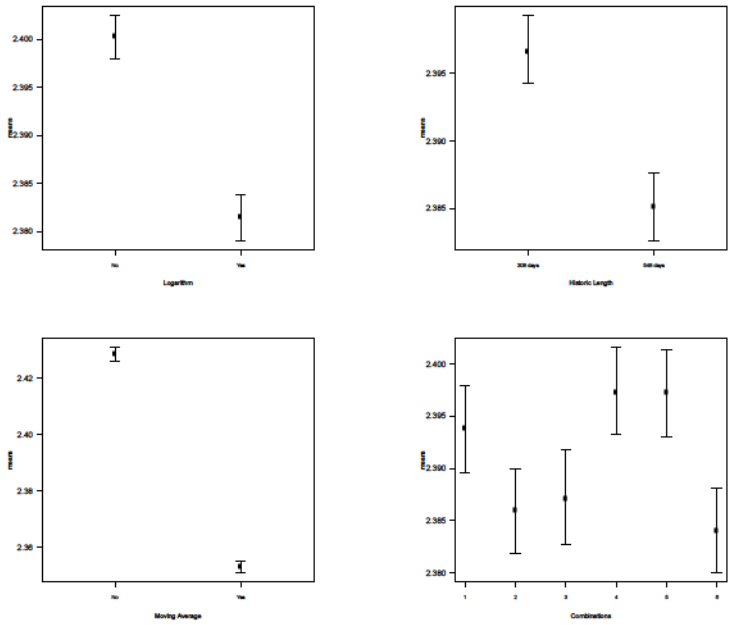


Figure B4. Bootstrap confidence intervals for the mean $ln(MAE)$ of the factors *Logarithm*, *Historical Length*, *Moving Average*, and *Forecast Combinations*. *Forecast Combinations* include: (1) benchmark BIC-selected model, (2) median-based combination, (3) mean-based combination, (4) BIC-based combination, (5) BIC-50% combination, (6) mean BIC-based combination. Forecast horizon $h = 60$, Italian market.

Appendix B.1.4. Minimizing Forecasting Error for Two-Month-Ahead Forecasts ($h = 60$)

For the largest forecasting horizon considered ($h = 60$) we obtain that all the factors included in the DOE are statistically significant (see Table B5). Figure B4 presents the confidence intervals showing the effect of the factors in the mean of $ln(MAE)$. Better results are obtained when the dependent variable is $ln(Prices)$ instead of *Prices*, and also when the long *Historical Length* is used, contrary to what was found in the previous forecasting horizons but supported in the literature. Setting *Moving Average*='Yes' still improves the results. We continue to find that *Forecast Combination* 6 provides significantly better results than the benchmark (*Combination* 1 in the figure), but the benefits of using 2 and 3 are not as strong as with $h < 60$.

Appendix C. ANOVA Bootstrap Procedures

Davison and Hinkley (1997) indicate that when we have doubts about the accuracy of the normal approximation, the empirical distribution can be a fairer approximation of the distribution of the parameter of interest. As an advantage, we do not need to know the distribution of the underlying parameter in order to employ bootstrap procedures.

Appendix C.1. Bootstrap Procedure to Calculate ANOVA P-values

The ANOVA performed relies on the assumption that the residuals u_{ijkld} of expression (11) are normally distributed. When the data is normally distributed, the Sum of Squares has a χ^2 distribution, and the quotient of Mean Squares follows a distribution Fisher-Snedecor (we call this statistic *F*-ratio). The lack of normality in u_{ijkld} results in an *F*-ratio that will most likely not have a Fisher-Snedecor distribution; therefore, the p-value, determined as $Pr(F-\text{ratio} \geq F_{n_1,n_2,\alpha})$[6], is no longer accurate. To avoid complicating notation, we keep using the denomination *F*-ratio even if this statistic does not have a Fisher-Snedecor distribution.

To be robust to departures from the Gaussianity assumption, we calculate the ANOVA p-values employing bootstrap, according to the following steps. We use the DOE factor *Logarithm* for illustrative purposes, but the procedure is the same for all the factors considered.

1. We estimate model (11), obtaining estimates $\widehat{u}_{ijkld}$ and the *F*-ratio of the ANOVA.
2. We would like to test if there is an effect to taking logarithm of prices. In other words, is there a difference in the forecasting error of setting *Logarithm* = *No* vs. *Logarithm* = *Yes*? This translates in the null hypothesis $H0: \alpha_{No} = \alpha_{Yes}$. We estimate (11) under null hypothesis $H0$ and obtain estimates μ^R, β_j^R, γ_k^R, δ_l^R and ϵ_d^R, where ^{R} stands for 'restricted' model.
3. We generate synthetic samples for the ln(MAE); each bootstrap replication needs to satisfy the null hypothesis, though employing a random sample of the unrestricted model's estimated residuals, $\widehat{u^*}_{ijkld}$. The replications are independent, and the random samples (with replacement) of $\widehat{u^*}_{ijkld}$ have the same size as the original data-set.

$$ln(MAE^*_{ijkld}) = \mu^R + \beta_j^R + \gamma_k^R + \delta_l^R + \epsilon_d^R + \widehat{u^*}_{ijkld}. \tag{12}$$

4. For each bootstrap replication we re-estimate model (11) to the synthetic data $ln(MAE^*_{ijkld})$ and save the statistic F_b^*-ratio, where the sub-index $b = 1, ..., B$ represents each of the B bootstrap replicas.
5. We obtain the Monte Carlo p-value as indicated in Davison and Hinkley (1997),

$$\hat{p}^* = \frac{1 + \#\{F_b^* \geq F\}}{B+1}, \tag{13}$$

where $\#\{\cdot\}$ indicates the number of times the event in braces occurs. We employ $B = 500$ bootstrap replications.

Appendix C.2. Bootstrap Procedure to Calculate ANOVA Main Effects Confidence Intervals

Also to permit departures from Gaussianity, the confidence intervals for the main effects need not be estimated with the parametric formula employed for normal data. Instead, we recur to a Monte Carlo simulation of the bootstrap. The procedure is the following.

1. We estimate model (11), obtaining estimates for the residuals, $\widehat{u}_{ijkld}$.
2. For each bootstrap replication, we obtain a random sample of $\widehat{u}_{ijkld}$, and generate a data-set

$$ln(MAE^*_{ijkld}) = \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + \epsilon_d + \widehat{u}^*_{ijkld}, \tag{14}$$

3. We use each simulated sample to estimate the ANOVA and obtain the estimates for each value of the factors *Logarithm*, *Historical Length*, *MA*, and *Forecast Combinations*.

---

[6] $F_{n_1,n_2,\alpha}$ is such that $Pr(F > F_{n1,n2,\alpha}) = \alpha$ when F follows the Fisher-Snedecor distribution with $n1$ and $n2$ degrees of freedom.

4. We work with $B = 500$ replications to obtain a bootstrap distribution for the mean ln(MAE) at each level of the factors of the DOE. We use the bootstrap percentile interval (Davison and Hinkley, 1997, chapter 5) to calculate 95% confidence intervals for the mean levels of each factor, by employing the 2.5 and 97.5 percentiles of the estimates of all the bootstrap replications.

Appendix D. Combinations and Benchmark Comparison Employing the RMSE

In this section, we present the results analogous to Tables 3 and 4, but employing a different accuracy metric, the Root Mean Squared Error. We obtain similar results to those for the MAE: for both data-sets, *Forecast Combinations* median-based, mean-based, and specially mean-BIC-based, report noticeable improvements with respect to the benchmark BIC-selected model.

Table D1. Weekly, monthly, and bi-monthly RMSE for the Iberian Market.

	BIC-selected model	Median-based Combination	Mean-based Combination	BIC-based Combination	BIC 50% Combination	Mean-BIC-based Combination
Weekly RMSE	7.1468	7.0810	7.1103	7.1389	7.1389	7.0393
Monthly RMSE	8.1668	7.9612	7.9721	8.1533	8.1533	7.9052
Bi-Monthly RMSE	9.1352	8.8689	8.8705	9.1189	9.1189	8.7971

Table D2. Weekly, monthly, and bi-monthly RMSE for the Italian Market.

	BIC-selected model	Median-based Combination	Mean-based Combination	BIC-based Combination	BIC 50% Combination	Mean-BIC-based Combination
Weekly RMSE	10.6280	10.5466	10.5937	10.6194	10.6194	10.4675
Monthly RMSE	11.9479	11.8175	11.8685	11.9353	11.9353	11.7096
Bi-Monthly RMSE	12.9861	12.6656	12.6981	12.9637	12.9637	12.5709

Bibliography

A. Alonso, C. García-Martos, J. Rodríguez, and M. Sánchez. Seasonal dynamic factor analysis and bootstrap inference: Application to electricity market forecasting. *Technometrics*, 53:137–151, 2011.

J. Bai and S. Ng. Determining the number of factors in approximate factor models. *Econometrica*, 70:191–221, 2002.

A. Bello, J. Reneses, and A. Muñoz. Medium-term probabilistic forecasting of extremely low prices in electricity markets: Application to the Spanish case. *Energies*, 9:193–219, 2016.

M. Billio, R. Casarin, F. Ravazzolo, and H. Van Dijk. Combining predictive densities using nonlinear filtering with applications to us economics data. *Tinbergen Institute Discussion Paper*, 2011. Available online: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1967435 (accessed on 18 July 2016).

H. Bjørnland, K. Gerdrup, A. Jore, C. Smith, and L. Thorsrud. Does forecast combination improve norges bank inflation forecasts? *CAMAR Working Paper Series*, 2:1–32, 2010.

S. Bordignon, D. Bunn, F. Lisi, and F. Nan. Combining day-ahead forecasts for British electricity prices. *Energy Economics*, 35:88–103, 2013.

H. Chipman, E. George, and R. McCulloch. The practical implementation of bayesian model selection. *IMS Lecture Notes*, 38:70–134, 2001.

R. Clemen. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5:559–583, 1989.

A. Conejo, J. Contreras, R. Espínola, and M. Plazas. Forecasting electricity prices for a day-ahead poll-based electric energy market. *International Journal of Forecasting*, 21:435–462, 2005.

K. Cremers. Stock return predictability: A Bayesian model selection perspective. *Review of Financial Studies*, 15:1223–1249, 2002.

A. C. Davison and D. R. Hinkley. *Bootstrap Methods and their Application*. Cambridge University Press., 1997.

V. Di Cosmo. Forward price, renewables and the electricity price: The case of Italy. *The Economic and Social Research Institute Working Paper*, 511:1–24, 2015. Available online: https://www.esri.ie/publications/forward-price-renewables-and-the-electricity-price-the-case-of-italy/ (accessed on 3 November 2016).

G. Elliott, C. Granger, and A. Timmermann. *Handbook of Economic Forecasting vol.1*, Oxford: North-Holland. edition, 2006.

C. García-Martos, J. Rodríguez, and M. Sánchez. Mixed models for short-run forecasting of electricity prices: Application for the Spanish market. *IEEE Transactions on Power Systems*, 22:544–552, 2007.

C. García-Martos, J. Rodríguez, and M. Sánchez. Forecasting electricity prices by extracting dynamic common factors: application to the Iberian market. *IET Generation, Transmission & Distribution*, 6:11–20, 2012.

C. García-Martos, E. Caro, and M. Sánchez. Electricity price forecasting accounting for renewable energies: optimal combined forecasts. *Journal of the Operational Research Society*, 66:871–884, 2015.

J. Geweke. *The dynamic factor analysis of economic time series. Latent variables in socio-economic models*. Amsterdam, North Holland, 1977.

R. Huisman, C. Huurman, and R. Mahieu. Hourly electricity prices in day-ahead markets. *Energy Economics*, 29:240–248, 2007.

R. Hyndman and A. Koehler. Another look at measures of accuracy. *International Journal of Forecasting*, 22: 679–688, 2006.

G. Koop and S. Potter. Forecasting in large macroeconomic panels using bayesian model averaging. *Federal Reserve Bank of New York, Staff Report*, 163, 2003.

V. Kuzin, M. Marcellino, and C. Schumacher. Pooling versus model selection for nowcasting GDP with many predictors: Empirical evidence for six industrialized countries. *Journal of Applied Econometrics*, 28: 32–411, 2012.

E. Leamer. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*. New York: John Wiley & Sons., 1978.

R. Lee and L. Carter. Modelling and forecasting US mortality. *Journal of the American Statistical Association*, 87:659–671, 1992.

K. Maciejowska and R. Weron. Forecasting of daily electricity prices with factor models: utilizing intra-day and inter-zone relationships. *Comput Stat*, 30:805–819, 2015.

F. Martínez-Álvarez, A. Troncoso, G. Asencio-Cortés, and J. Riquelme. A survey on data mining techniques applied to electricity-related time series forecasting. *Energies*, 8:13162—-13193, 2015.

C. Monteiro, L. Fernandez-Jimenez, and I. Ramirez-Rosado. Explanatory information analysis for day-ahead price forecasting in the Iberian electricity market. *Energies*, 8:10464–10486, 2015.

D. C. Montgomery. *Design and Analysis of Experiments*. New York: Wiley, 1984.

J. Nowotarskia, E. Raviv, S. Trückc, and R. Werona. An empirical comparison of alternate schemes for combining electricity spot price forecasts. *Energy Economics*, 46:395–412, 2014.

A. Panagiotelis and M. Smith. Bayesian density forecasting of intraday electricity prices using multivariate skewed t distributions. *International Journal of Forecasting*, 24:710–727, 2008.

D. Peña and G. Box. Identifying a simplifying structure in time series. *Journal of the American Statistical Association*, 82:836–843, 1987.

D. Peña and P. Poncela. Forecasting with nonstationary dynamic factor models. *Journal of Econometrics*, 119: 291–321, 2004.

D. Peña and P. Poncela. Nonstationary dynamic factor analysis. *Journal of Statistical Planning and Inference*, 136:1237–1257, 2006.

P. Poncela, J. Rodriguez, R. Sánchez-Mangas, and E. Senra. Forecast combination through dimension reduction techniques. *Journal of Forecasting*, 27:224–237, 2011.

A. Raftery. Bayesian model selection in social research. *Sociological Methodology*, 25:111–163, 1995.

A. Raftery, D. Madigan, and J. Hoeting. Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92:179–191, 1997.

E. Raviv, K. Bouwman, and D. Van Dijk. Forecasting day-ahead electricity prices: utilizing hourly prices. *Energy Economics*, 50:227—239, 2015.

I. Sánchez. Short-term prediction of wind energy production. *International Journal of Forecasting*, 22:43–56, 2006.

T. Sargent and C. Sims. Business cycle modelling without pretending to have too much a priori economic theory,new methods in business cycle research. *Federal Reserve Bank of Minneapolis Working papers*, 55, 1977.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464, 1978.

J. Stock and M. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97:1167–1179, 2002.

J. Stock and M. Watson. Combination forecasts of output growth in a seven country dataset. *Journal of Forecasting*, 23:405–430, 2004.

N. Swanson and T. Zeng. Choosing among competing econometric forecasts: Regression-based forecast combination using model selection. *Journal of Forecasting*, 20:425–440, 2001.

R. Weron. Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 30:1030—-1081, 2014.

J. Wright. Bayesian model averaging and exchange rate forecasts. *Journal of Econometrics*, 146:329–341, 2008.

Y. Yang. Combining forecasting procedures: some theoretical results. *Econometric Theory*, 20:176–222, 2004.