International Statistical Review
Revue Internationale de Statistique

International Statistical Review (2013), 81, 2, 289–306 doi:10.1111/insr.12014

Improving Electricity Market Price Forecasting with Factor Models for the Optimal Generation Bid

M. Pilar Muñoz[1], Cristina Corchero[1,2] and F.-Javier Heredia[1]

[1] *Department of Statistics and Operations Research, Universitat Politecnica de Catalunya, Jordi Girona 1-3, Office 205, Barcelona, 08034, Spain*
E-mails: pilar.munyoz@upc.edu, f.javier.heredia@upc.edu
[2] *Catalonia Institute for Energy Research, Jardins de les Dones de Negre 1, 2nd floor, Sant Adrià del Besos, 08930, Spain*
E-mail: cristina.corchero@upc.edu

Summary

In liberalized electricity markets, the electricity generation companies usually manage their production by developing hourly bids that are sent to the day-ahead market. As the prices at which the energy will be purchased are unknown until the end of the bidding process, forecasting of spot prices has become an essential element in electricity management strategies. In this article, we apply forecasting factor models to the market framework in Spain and Portugal and study their performance. Although their goodness of fit is similar to that of autoregressive integrated moving average models, they are easier to implement. The second part of the paper uses the spot-price forecasting model to generate inputs for a stochastic programming model, which is then used to determine the company's optimal generation bid. The resulting optimal bidding curves are presented and analyzed in the context of the Iberian day-ahead electricity market.

Key words: Electricity market prices; short-term forecasting; stochastic programming; factor models; price scenarios.

1 Introduction

Revenues of electricity generation companies (GenCos) were traditionally determined by a wholesale price that was established by either its customers (electricity distribution companies) using bilateral contracts or by the government (in the case of domestic customers). The creation of the electricity trading markets has liberalized the electricity sector in Spain and Portugal (with the creation of the Iberian Energy Market) and elsewhere, leading to uncertainty in the price. In this new environment, the GenCos must develop and submit their bids to the market without knowing the final prices that will be paid for the electricity they produce. As a result, liberalized electricity markets around the world need to develop good forecasts of the prices at which the energy will be paid in order to decide on the bids and how to schedule their resources for maximizing their profits (Shahidehpour *et al.,* 2002).

The objectives of this paper are twofold. The first part deals with methods for short-term forecasting electricity market spot prices. The second part describes how they can be used as

inputs into a stochastic programming optimization framework to develop the bids (Birge & Louveaux, 1997). We apply the results to the Iberian Electricity Market. As far as we know, this is the first attempt to build a stochastic programming model for the optimal day-ahead bidding in electricity markets by using price scenarios based on time-series factor analysis (TSFA). This optimal day-ahead bidding is an important problem for the GenCos. The simplicity of the TSFA, in comparison with the alternative methods, is an important practical contribution of this paper.

Electricity spot prices exhibit many types of non-stationary behavior: non-constant mean and variance, daily and weekly seasonality, calendar effects due to weekends and holidays, and high volatility. These characteristics make it challenging to develop reliable short-term forecasts of electricity prices. A review of various models and methods is provided in the next section. All of these have advantages and drawbacks, but they are quite adequate for the purposes of modeling prices at hourly levels. In this paper, we use instead the well-known factor models to forecast electricity market prices on a short-term horizon (24 h). A review of these methods is also provided in the next section. The performance of these methods is similar to that of the autoregressive integrated moving average (ARIMA)-based methods, but they are easier to implement. In applying the factor models, the spot prices are interpreted as a set of 24 different time series, one for each hour of the day, in a manner similar to that of Munoz & Bunk (2007), Alonso *et al*. (2008), and Karakatsani & Bunn (2008). The factor models allow us to characterize common unobserved factors that represent the relationships between the hours of a day. Dynamic and static factor models have been extensively used in many different contexts (Geweke, 1977; Stock & Watson, 2002; Peña & Poncela, 2004 or Peña & Poncela, 2006). However, they have not been used for forecasting short-term electricity market prices. In this paper, we use static factor models to develop and evaluate spot prices for the Iberian Electricity Market. Previous results have shown that dynamic factor models are better (Forni *et al.*, 2005; Doz *et al.*, 2012), but some authors (Stock & Watson, 2002; Boivin & Ng, 2005) have noted that the benefits are not big enough to compensate for the ease of use of and interpretation of static factor models.

The second part of the paper uses the forecasted prices as inputs for a stochastic programming forecasting model. Stochastic programming (Birge & Louveaux, 1997) is an effective methodology that allows stochastic aspects to be incorporated into traditional linear, integer, and nonlinear optimization problems through the definition of a scenario tree (Ruszczynski & Shapiro, 2003; Wallace & Fleten, 20032003). This optimization technique is designed to introduce the uncertainty into the model via the scenario tree. To introduce the mathematical form of a distribution into a mathematical programming model is usually not viable, and the most common approach is to build a set of scenarios, which approximate the empirical distribution.

The optimization problem will focus on the day-ahead electricity market bid of a GenCo taking into account the committed physical derivatives products, following the approach in Corchero & Heredia (2011). Other approaches to such problems involving future contracts can be found in Chen *et al*. (2004) and, for medium-term optimization, Conejo *et al*. (2008) and Guan *et al*. (2008). In our case, the stochastic variable is the day-ahead market clearing price. So, a set of scenarios for the day-ahead market clearing price will be built using the forecasting results. This set of scenarios is then introduced into the optimization problem, and stability analysis and results are obtained.

The rest of the article is organized as follows. Section 2 contains a brief introduction to the Iberian Energy Market and other energy concepts. Section 3 starts with a review of the literature and describes the factor model used. The stochastic optimization procedures are discussed in Section 4. The application of the results to the Iberian Electricity Market is described in Section 5. The paper ends with some concluding remarks.

2 Iberian Energy Market

In the present Iberian electricity system, the day-ahead market is the most important in terms of the electricity market with regard to physical energy exchanges. The objective of this market is to carry out the energy transactions for the next day by means of selling and buying offers presented by the market agents. The clearing process is based on the construction of an aggregated offer curve and the perceived demand curve for the 24 h of the next day. The offer curve consists of the buying bids sorted by increasing price so that the matched bids (those that will be actually settled) are the ones with lower prices. The first step of this curve is always the so-called zero-price bid, which consists of an amount of energy that is offered to the market for free (typically green resources energy that must be produced by the GenCo and the quantity coming from the futures contracts, as will be explained). On the other hand, the demand curve consists of the selling bids sorted by decreasing price so that the matched bids (those that will be actually settled) are the ones with higher prices.

The spot price (which is basically a settlement price paid), λ, is determined by the intersection of the perceived aggregate offer and demand curves (Figure 1), that is, by the equilibrium point. The resulting quantity p corresponds to the power that will be produced in the energy system in this hour.

The *physical futures contracts* are actually financial derivatives with actual physical delivery of the product. They are traded via organized derivatives markets. The main characteristics of these products are as follows:

- Procurement: They have cash settlement and physical delivery. The financial settlement corresponds to the difference between the day-ahead price and the futures reference price.
- Delivery period: The duration of the contract could be a year, a quarter, a month, or a week. During all the days within the delivery period, the quantity determined must be produced.
- Load: In the Iberian Electricity Market, all the contracts are base load, which means that they determine a constant quantity for all the hours of the day.

When a GenCo has a *physical futures contract* to settle, the market operator demands that this GenCo commits the quantity designated to the contract through the day-ahead market bid of the physical units. This commitment must be made by what is called a *price acceptant sale bid*, which is a sale bid that will be accepted regardless of the clearing price, that is with a price

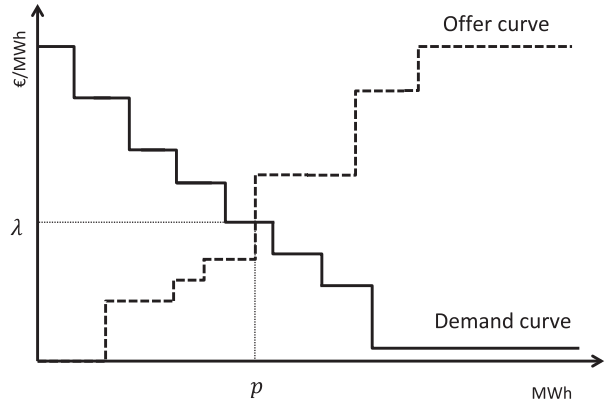

Figure 1. *Typical offer and demand curves for a certain hour.*

of 0€/kWh. That regulation implies that the GenCo has to determine its optimal bid by taking into account those price acceptant sale bids. Because of the algorithm that the market operator uses to clear the day-ahead market, all price acceptant sale bids will be matched (i.e., accepted) in the clearing process; that is, this energy shall be produced and will be remunerated at the spot price λ€/kWh.

3 Factor Models

3.1 Literature Review

Several approaches have been proposed in the power system literature for developing reliable short-term forecasts of electricity prices, and they can be basically classified into parametric or nonparametric and conditional homoscedastic or heteroscedastic. These approaches range from the well-known ARIMA models, which are part of the class of parametric-conditional homoscedastic models, to more complex nonparametric ones, as for example linear regression trees (Breiman *et al.*, 1984), kernel regression (Browman & Azzalini, 1997; Härdle, 1990), generalized additive models (Hastie & Tibsbshirani, 1990), wavelets (Stevenson, 2001), and artificial neural networks (Wang & Ramsey, 1998), among others. For a detailed description of the nonparametric methodology applied to forecasting electricity prices, see Mendes *et al.* (2008) and the references therein. Baillo *et al.* (2006) used clustering techniques on historic series of data to obtain possible price scenarios along with certain probability distributions. Contreras *et al.* (2003) used ARIMA models for forecasting day-ahead electricity prices. Conejo *et al.* (2005) compared the performances of day-ahead electricity price forecasting on the basis of ARIMA models, dynamic regression, transfer functions, wavelet-transforms, and neural networks for the Pennsylvania, Jersey, and Maryland market. For their data, predictions from dynamic regression and transfer function procedures were better than those from ARIMA models. Wavelets-based models had results close to ARIMA models, but neural network algorithms did not do well. Garcia-Martos *et al.* (2007) decomposed the Spanish hourly electricity prices into 24 individual time series and analyzed them separately, obtaining 1-day-ahead forecasts for each series. However, in most cases, the residuals exhibited a non-stationary conditional variance. This problem can be addressed by the classical generalized autoregressive conditional heteroscedasticity (GARCH) models and their variants to handle the conditional heteroscedasticity of electricity spot prices. Garcia *et al.* (2005) used an autoregressive moving average model with GARCH errors for the Spanish and California Electricity markets and showed that this combined approach overcomes the problems faced by the ARIMA model. Koopman *et al.* (2007) extended the approach to periodic dynamic long-memory regression models with GARCH errors.

As noted earlier, we will focus on factor models in this paper. Methods for estimation and forecasting of multivariate time series by using factor models can be classified into two main categories: static and dynamic. The static models are characterized by the fact that the factors and the associated weight matrix are constant over time, whereas the dynamic models allow latent variables, which are unobservable, to evolve dynamically over time.

The work of Cattell *et al.* (1947) is one of the earliest papers in static factor analysis, and it applied a principal component analysis to a set of multivariate time series. This procedure, called the *P-technique* by Cattell, was criticized by others mainly because it did not clearly specify the underlying assumptions (see Molenaar (1985) and the references therein). In addition, Cattell (1963) himself noted that the procedure did not take into account the lagged relationships between factors and the variables under study. McCallum (1970) proposed an alternative

that defined principal components estimators with minimum variances by using regression but noted that these estimators can be biased. Many authors have applied static factor analysis to reduce the dimensionality of the set of multivariate series. For example, Stock & Watson (2002) forecasted the values of a time series by a using principal component analysis; Haan *et al.* (2003) applied a static factor model to assess the quality of various economic indicators; Bai & Ng (2002) and Bai (2003) developed inferential theory when the number of series and the number of their dimension are both large. Pan & Yao (2008) developed a factor model for non-stationary time series and also provided its asymptotic properties. Their proposed methodology consists of assuming that there is no linear combination of factors that are white noise; otherwise, the random error space should be expanded with these linear combinations. The factor loadings are estimated by means of a sequence of nonlinear stepwise optimization algorithms.

In terms of dynamic factor models, some of the pioneering work on estimating factor models was written by Geweke (1977) and Sargent & Sims (1977), both on frequency domain, in other words, using the Fourier transform of the autocovariance function. Some of the early papers that studied dynamic factor models in the time domain were of Engle & Watson (1981) and Peña & Box (1987). Watson & Engle (1983) formulated the problem of dynamic factor analysis by using state-space representation and estimated the parameters by using the Kalman filter via the expectation–maximization (EM) algorithm (Shumway & Stoffer, 1982). The development of dynamic factor models is still rapidly evolving, as evidenced by recent papers (see Jungbacker *et al.* (2011), Proietti (2011), and references therein).

There is still considerable discussion of the relative merits of static and dynamic models. Bai & Ng (2008) noted that estimating the factors by using principal components is computationally simple and fast; however, they used only a simulation study to detect whether the factor structure has problems when the data set has a weak factor structure. The authors pointed out that the estimation of the factors seems to work well for low cross-sectional correlations, although the estimated factors are very sensitive to data sets with low signal-to-noise ratio in the sense that low ratios indicate a weak factor structure. Also, the problem of cross-sectional error cannot be treated with principal components. Schumacher (2007) used both static and dynamic factor models for forecasting German gross domestic product (GDP). The dynamic model outperformed the forecast accuracy of the static one, although the differences between their mean square forecast error are not very big. Marcelino & Schumacher (2010) proposed a different method based on mixed-frequency data sampling for forecasting the German GDP. Parameter estimation was carried out using three different procedures on the basis of dynamic, static, and Kalman filter algorithms. The nowcast performance was not influenced very much by the choice of the factor estimation procedure previously stated, and the factor estimation methods do not differ very much.

Time-series factor analysis is an alternative to both static and dynamic factor analysis methods. TSFA makes fewer assumptions compared with the dynamic method. For example, it does not assume covariance stationarity. The methodology proposed in TSFA follows the procedure proposed by Spanos (1984). See Gilbert & Meijer (2005) for an illustrative example of how TSFA works. Code for analyzing the data by using TSFA is available in the R package *tsfa* available on CRAN (http://cran.r-project.org/web/packages/tsfa/index.html). There are, however, complex situations (for example, when the sets of the time series are at different frequencies such as monthly and quarterly), where TSFA cannot be used. In these cases, Proietti (2011) proposed a new procedure, based on DFA, which allows working with series that have different frequencies. The model is formulated in classical state-space form, in which the transition equation describes the evolution of the unobserved factors. In this case, they follow an autoregressive process of order 1 while the time series are linearly related with the unobserved factor via the measurement equation. The parameter estimation can be performed by

maximizing the likelihood either through combining the Kalman filter and a quasi-Newton type of numerical optimization algorithm (Nocedal & Wright, 2000) or by using the EM algorithm proposed by Dempster *et al.* (1977). It is obvious that this procedure is much more flexible than the TSFA, but also much more complex because of the problems of computation, convergence, and misspecification that can occur.

3.2 Factor Model Estimation

Let y_t be an M-vector of an observed time series of length T and n be the number of unobserved factors or latent variables ($n << M$) collected in the n-vector ξ. The relationship between the observed time series y_t and the factors ξ is assumed to be linear and described by the equation

$$y_t = \alpha_t + B\xi_t + \epsilon_t, \tag{1}$$

where α_t is an M-vector of intercept parameters. If we work with centered data, α_t can be ignored. B is an $M \times n$ parameter matrix of loadings and is assumed to be time invariant, and ϵ is a random M-vector of errors. The random vector ϵ is assumed to be uncorrelated with the latent variables ξ.

We further assume that the model for y_t is integrated of order 1; that is, by taking the first difference, the data become stationary. Denoting D as the difference operator, (1) becomes

$$Dy_t \equiv y_t - y_{t-1} = (\alpha_t - \alpha_{t-1}) + B(\xi_t - \xi_{t-1}) + (\epsilon_t - \epsilon_{t-1}) \tag{2}$$

or

$$Dy_t = \tau_t + BD\xi_t + D\epsilon_t, \tag{3}$$

where $\tau_t = (\alpha_t - \alpha_{t-1})$. Further,

$$E\,D\xi_t = \kappa, \text{ the factor mean which exists and is finite} \tag{4}$$

$$E\left[(D\xi_t - \kappa)(D\xi_t - \kappa)'\right] = \Phi, \text{ the } n \times n \text{ factor variance–covariance matrix,} \tag{5}$$

which exists, has finite entries, and is positive definite

$$E\left[D\epsilon_t D\epsilon_t'\right] = \Omega, \text{ the } M \times M \text{ error variance–covariance matrix,} \tag{6}$$

which exists, has finite entries, and is positive definite

$$E(D\epsilon_t) = 0 \tag{7}$$

$$E\left[(D\xi_t - \kappa)D\epsilon_t'\right] = 0. \tag{8}$$

Define the sample mean and covariance of Dy_t as

$$\overline{Dy} \equiv \frac{1}{T}\sum_{t=1}^{T} Dy_t \tag{9}$$

$$S_{Dy} \equiv \frac{1}{T}\sum_{t=1}^{T}\left(Dy_t - \overline{Dy_t}\right)\left(Dy_t - \overline{Dy_t}\right)'. \tag{10}$$

Then it can be shown that

$$\overline{Dy} \xrightarrow{p} \mu \equiv \tau + B\kappa \tag{11}$$

$$S_{Dy} \xrightarrow{p} \Sigma \equiv B\Phi B' + \Omega. \tag{12}$$

Anderson & Amemiya (1988) discussed methods for estimating the parameters in $\theta = (B, \Phi, \Omega)$, where B is the loadings matrix, and Φ and Ω are the covariance matrix of the factors and errors, respectively. They maximized the Wishart likelihood on the basis of the empirical covariance matrix S_{Dy}. If the underlying data are normal, TS_{Dy} follows a Wishart distribution $W_n(\Sigma, T-1)$, so the likelihood is

$$L(\theta, S_{Dy}) \equiv \ln\left(|\Sigma(\theta)|\right) + \mathrm{tr}\left(S_{Dy}\Sigma^{-1}(\theta)\right), \tag{13}$$

where θ is the parameter vector to be estimated.

Once the parameters are estimated, the factor scores or the latent variables must be predicted. The best linear unbiased predictor of the factor scores is the Bartlett predictor (Wansbeek & Meijer, 2000), obtained from equation (14), assuming that the intercept $\tau_t = 0$:

$$\hat{\xi}_t = \left(\hat{B}'\hat{\Omega}\hat{B}\right)^{-1}\hat{B}'\hat{\Omega}^{-1}y_t, \tag{14}$$

where $\hat{B}$ and $\hat{\Omega}$ are the loadings and covariance error matrices, respectively, already computed in the factor analysis estimation step.

3.3 Determining the Number of Factors

Prior to the estimation of the parameters in the factor analysis procedure, the number of factors must be chosen. There are two methods for deciding the number of factors, one based on the likelihood ratio test (Tsay, 2010) and the other one associated with the spectral decomposition of the correlation matrix of the multivariate time series. We consider the second one in this paper. To determine a starting point for the number of factors, the following rule of thumb (Fabrigar *et al.*, 1999) is used: that the number of factors should be equal to the number of eigenvalues of the correlation matrix that are greater than 1. Another guideline that is also used is that the number of factors is the number of eigenvalues before the kink in the scree plot. The scree plot is the plot of eigenvalues (in descending order) of the correlation matrix versus their sequence number (Wansbeek & Meijer, 2000). If the points on the plot are connected sequentially, then the place where meaningful change in the slope takes place is called the kink, and it almost always exists in practice.

3.4 Forecasting and Scenario Generation

We can now use the predicted factors from Section 3.2 in a forecasting model to obtain the price forecasts. For example, the one-step-ahead forecasting model is specified as a linear regression model

$$y_{t+1} = \beta\hat{\xi}_t + \alpha(L)y_t + \varepsilon_{t+1} \tag{15}$$

with the factors as predictors. Here $\hat{\xi}_t$ is the vector of predicted factors, β is the loadings matrix, and ε_{t+1} is the resulting forecast error. Autoregressive terms are included by using the polynomial of the non-negative power of the lag operator L with $\alpha(L) = \alpha_1 + \alpha_2 L + \alpha_3 L^2 + \ldots$.

The out-of-sample forecasts for y_{T+1} are conditional on the information observed until period T and is given by the conditional expectation

$$\hat{y}_{T+1|T} = \hat{\beta}\hat{\xi}_T + \hat{\alpha}(L)y_T. \tag{16}$$

As discussed earlier, our objective is to obtain representative scenarios—a set of possible values for the electricity prices for time t, with $t = 1$ to $t = 24$, or in general $t = T$, $\lambda_t^s = \{\lambda_1^s, \ldots, \lambda_T^s\}$ and their corresponding probabilities P^s for all the sets of scenarios. The problem of building the scenario tree has been tackled by many authors (see Kaut & Wallace (2003) and Dupacova *et al.* (2000) for a survey). We will follow the next steps:

(1) Sampling: We use a procedure based on bootstrap techniques (Efron & Tibshirani, 1993) in order to obtain a set of scenarios. It is known that by increasing the number of scenarios, the empirical distribution function will approximate the theoretical one.
(2) Reduction: Given that the size and computational cost of the programming models depends on the number of scenarios, some scenario reduction techniques have to be applied in order to reduce the generated set of scenarios into one that is smaller but representative. We have applied the scenario reduction algorithm explained in Growe-Kuska *et al.* (2003); this proposed algorithm determines a subset of the initial scenario set and assigns new probabilities to the preserved scenarios. Construction of the scenario tree successively reduces the number of nodes in a fan of scenarios by modifying the tree structure and by bundling similar scenarios. The whole procedure is based on a recursive reduction argument using a transportation metric. The algorithms control the goodness of fit of the approximation through a probability metric.

Bootstrap techniques have been applied for estimating uncertainty in order to forecast and build confidence intervals (Alonso *et al.*, 2008). In the case of TSFA models, we use the following procedure that is heavily based on bootstrap:

(1) The model is estimated by means of maximum likelihood, and the estimates of the parameters are obtained based on the entire data set of independent and dependent variables of interest.
(2) The estimated residuals, $\hat{\varepsilon_t} = y_t - \left(\hat{\alpha}_t + \hat{B}\hat{\xi}_t\right)$, are obtained.
(3) An i.i.d. resample $\tilde{\varepsilon}_t$ from $F_{\hat{\varepsilon}_t}$, for $t = 1, \ldots, M$ is obtained, where $F_{\hat{\varepsilon}_t}$ is the empirical distribution function of the $\hat{\varepsilon}_t$.
(4) A bootstrap replica of the data, defined by $\tilde{y} = \hat{\alpha}_t + \hat{B}\hat{\xi}_t + \tilde{\varepsilon}_t$, is built.
(5) A sample of the future $\tilde{\varepsilon}_{t+h}$ is generated by resampling from $F_{\hat{\varepsilon}}$ H times (that is, for $h = 1, \ldots, H$, where H is the forecasting horizon).
(6) Future bootstrap observations are calculated using the equation in Step 4.

3.5 Results

This section discusses the results on the validation of the proposed forecasting method. The variable to be forecasted is the Iberian Electricity Market spot prices. These are hourly data for the work days from 1 January 2007 to 30 March 2008 (www.omel.es). As noted earlier, we treat the hourly electricity prices as a set of 24 time series, one for each hour. These 24 time series must be summarized by a small number of factors.

The number of factors, based on the eigenvalues of the sample correlation matrix and the kink in the scree plot, was determined to be three. See Figure 2 for the scree plot.

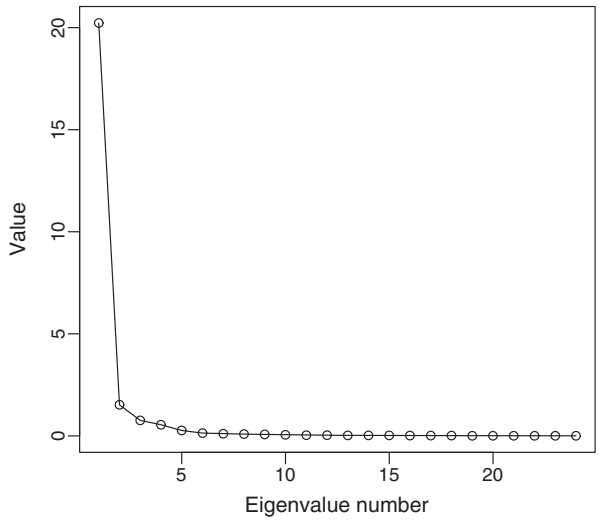

Figure 2. *Scree plot of the eigenvalues of the correlation matrix obtained from 24 time-series electricity prices.*

The obtained loadings matrix is represented in Figure 3 with h1 = 1:00 (assuming a 24-h clock), and the boxplots of hourly prices are shown in Figure 4. The behavior of the prices throughout 1 day has a particular profile, with hours called *base hours* and *peak hours*. During the base hours, the price is low and the variance is lower than in non-base hours, as can be seen in Figure 4. On the other hand, during the peak hours, the prices are the highest and they have high variance. A certain relationship can be observed between this profile and the factors loadings. The first factor, in Figure 3, clearly separates between night and day, and it can be observed that the profile of the daily hours' factor loadings (between 8:00 and 20:00 h) is similar to the profile of the prices during these hours. The second factor gives positive factor loadings to the base hours, and the third factor gives positive factor loadings to the peak hours.

Table 1 contains the factor loadings for the three-factor solution and the corresponding communality; Table 2 contains the goodness-of-fit statistics for the one-factor, two-factor, and three-factor solution (Wansbeek & Meijer, 2000):

- The comparative fit index (CFI) is a pseudo-R^2 statistic that is based on the χ-squared statistic and whose value is always between 0 and 1.
- Root mean square error of approximation (RMSEA) is a non-negative number, based also on the χ-squared statistic that measures the lack of fit per degree of freedom.
- Communality is the proportion of the variance of the original variable explained by each factor. In this case, there is a communality calculated for each of the 24 original hourly time series, and it represents the proportion of the variance of the electricity price at hour k, explained by the three common factors.

Table 1 reveals several interesting findings. The communalities shown in the last column indicate the relative importance of each variable with respect to its variance, in that those hours with higher communalities are better fit by the three factors than those with lower communalities. In this case, the highest communalities are for hours 3:00, 4:00, 5:00, and 6:00 as well as for the hours from 11:00 to 18:00. When inspecting the factor loadings within a row (which represents an independent variable in the data set that served as the data for the factor analysis),

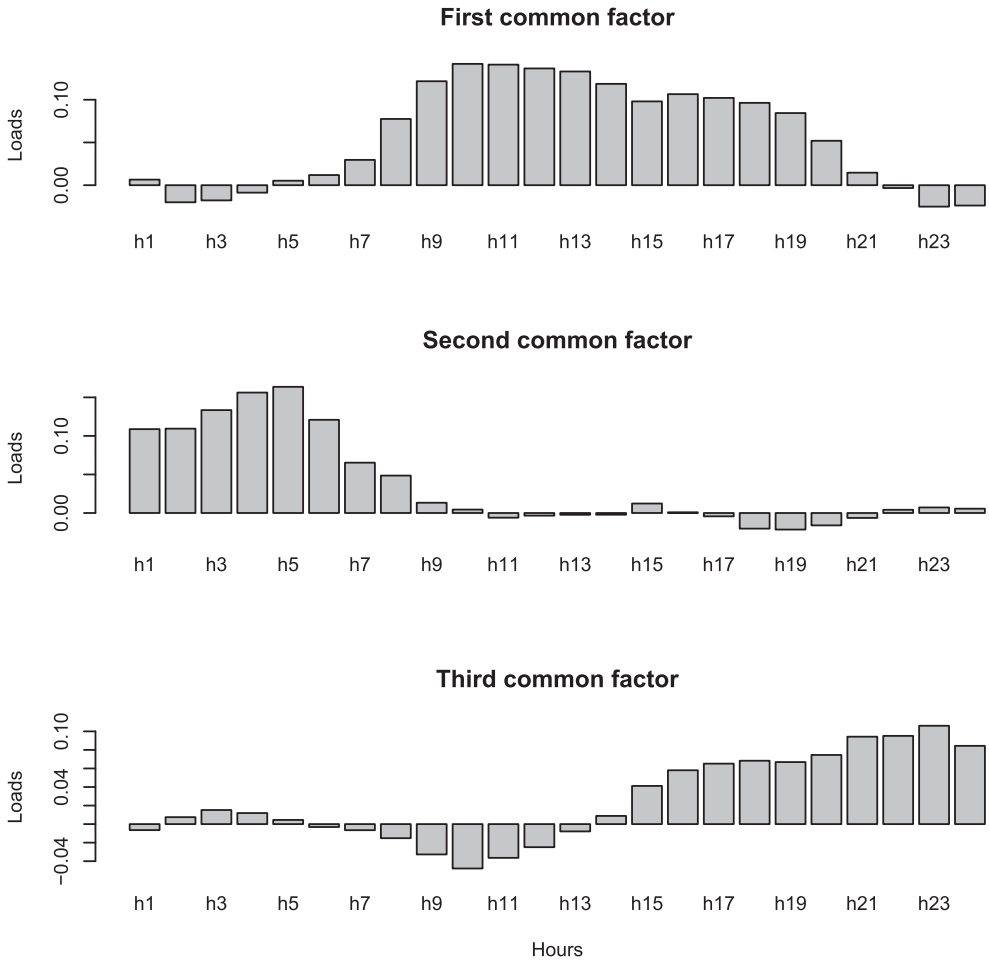

Figure 3. *Factor loadings for the first three factors.*

the factors with the highest magnitude loadings are the focus of the practical interpretation as to which factors are most related to those independent variables (Tsay, 2010). Factor 2 explains the behavior of the first group of hours mentioned (from 3:00 to 6:00), and Factor 1 is a good representative of the electricity prices for the hours from 11:00 to 18:00.

It can be seen in Table 2 that with three factors, we obtain a CFI equal to 0.755 and an RMSEA equal to 0.190, showing that three factors are enough for obtaining a good representation of the price evolution during a 24-h period. The results in Table 2 show that increasing the number of factors up to 4 does not improve the goodness-of-fit measurements enough to justify this new factor. We must increase the number of factors to 10 to obtain significant changes in those goodness-of-fit measurements over the three-factor solution.

The forecasting model is thus based on these three factors. The estimation of the 24 regression models based on the three factors is made with the available data set, keeping the data of the last week available aside, in order to check the forecast regression model. The R^2 and the mean square error (MSE) of the forecast regression model for each hour are shown in Table 3. The estimated models are used to forecast the next 5 days. Figure 5 plots the actual price (light line), the forecast price (dark line), and the forecast set of bootstrap replicates (gray lines) used

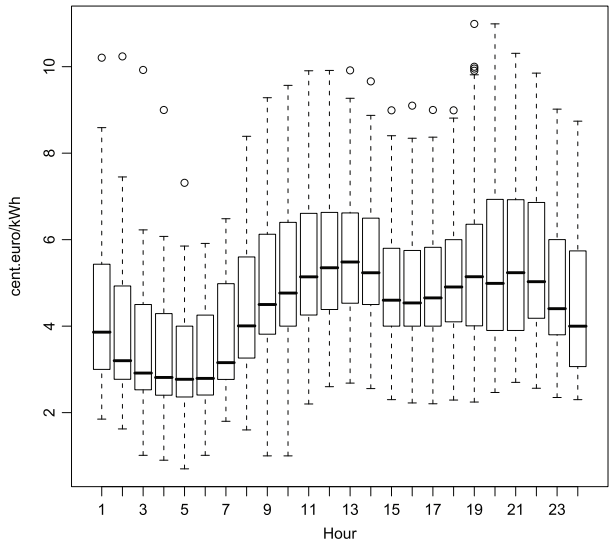

Figure 4. *Iberian Electricity Market spot prices by hour for the period of 1 January 2007 to 30 March 2008.*

Table 1. *Factors loadings and communality for each hour.*

Hour	Factor 1	Factor 2	Factor 3	Communality
1	0.007	0.109	−0.006	0.419
2	−0.020	0.109	0.007	0.669
3	−0.018	0.133	0.015	0.844
4	−0.009	0.156	0.011	0.944
5	0.005	0.164	0.005	0.911
6	0.011	0.121	−0.003	0.818
7	0.030	0.065	−0.006	0.547
8	0.078	0.049	−0.015	0.461
9	0.122	0.013	−0.033	0.585
10	0.142	0.004	−0.048	0.737
11	0.141	−0.006	−0.036	0.903
12	0.137	−0.003	−0.025	0.932
13	0.133	−0.002	−0.008	0.895
14	0.119	−0.002	0.008	0.839
15	0.098	0.012	0.041	0.730
16	0.106	0.001	0.058	0.795
17	0.102	−0.004	0.065	0.822
18	0.096	−0.020	0.068	0.832
19	0.084	−0.021	0.067	0.719
20	0.052	−0.016	0.075	0.633
21	0.015	−0.006	0.094	0.557
22	−0.003	0.004	0.095	0.608
23	−0.025	0.007	0.106	0.561
24	−0.024	0.006	0.084	0.366

to build the set of scenarios. This forecasting procedure has been compared with the ARIMA model used in previous works (Corchero, 2010), and it has been observed that the results in terms of MSE are equivalent.

Table 2. *Measures of goodness of fit for the one-factor, two-factor, and three-factor solutions.*

Statistic	1 Factor model	2 Factor model	3 Factor model	4 Factor model	... ...	10 Factor model
CFI	0.347	0.587	0.755	0.834	...	0.990
RMSEA	0.289	0.241	0.190	0.174	...	0.058

Table 3. *Summary for the forecast models for each hour.*

Hour	1	2	3	4	5	6	7	8	9	10	11	12
R^2	99.1	95.3	97.1	99.8	99.8	97.6	96.0	99.6	99.7	99.8	96.3	98.3
MSE	0.017	0.004	0.003	0.003	0.002	0.002	0.003	0.008	0.008	0.004	0.003	0.001
Hour	13	14	15	16	17	18	19	20	21	22	23	24
R^2	99.9	97.7	99.8	99.9	99.9	97.1	99.7	96.6	94.2	99.7	99.7	95.1
MSE	0.002	0.002	0.004	0.002	0.002	0.002	0.006	0.005	0.007	0.007	0.007	0.005

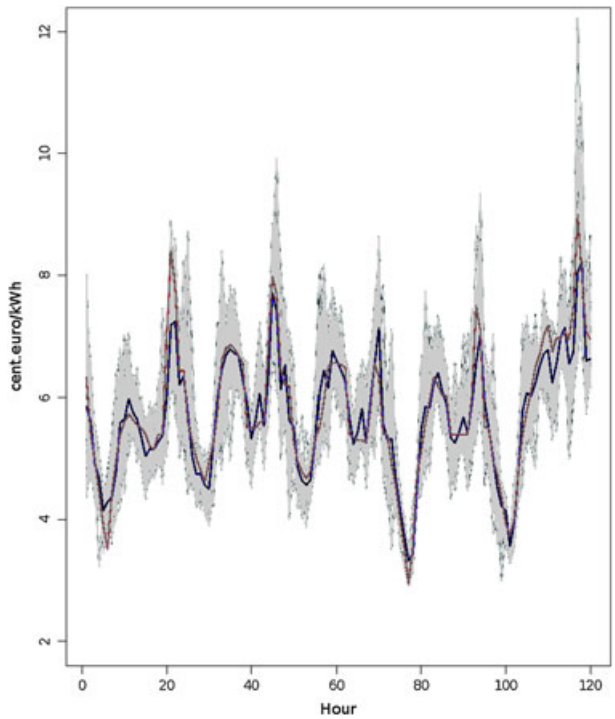

Figure 5. *One-step-ahead forecast prices and bootstrap replicates.*

Those forecasts must be updated once the new prices are known. This is performed using the same model and estimated parameters with the new prices in order to build the new forecast.

4 Application to Energy Market Optimization

4.1 Stochastic Optimization Model

The optimization model to be discussed in this section for a single GenCo has the following parameters:

- Sets:
  - I, the set of thermal units of a particular GenCo participating in the auction process;
  - F, the set of physical futures contracts of the GenCo;
  - S, the set of day-ahead price scenarios;
  - F_i, the subset of contracts in which unit $i \in I$ is involved;
  - I_j, the set of thermal units assigned to contract j.
- Parameters:
  - c_i^b (€), c_i^l (€/kWh), and c_i^q (€/kWh^2), coefficients (constant, linear, and quadratic coefficients, respectively) for the quadratic generation cost function for unit i;
  - $\overline{P}_i$ and $\underline{P}_i$, the upper and lower bounds on the energy generation (kWh) for unit i, respectively;
  - λ^s day-ahead price scenario s;
  - L_j quantity to be settled in physical futures contract j;
  - P^s probability of scenario s.
- Variables:
  - q_i, the quantity of energy involved in the price acceptant sale bids, that is, the quantity of energy bid by unit i to the auction of the day-ahead market at 0€/kWh;
  - f_{ij}, the energy delivered by the thermal unit i to the *physical futures contract* j;
  - p_i^s, the matched energy of thermal i at the auction of the day-ahead market under scenario s.

The problem is to maximize the expected value of the benefits coming from the day-ahead market. At each scenario $s \in S$, this value can be calculated as the difference between the incomes from the matched energy, $\lambda^s p_i^s$, and the quadratic generations costs $\left(c_i^b + c_i^l p_i^s + c_i^q (p_i^s)^2\right)$ (Zhu, 2009). This maximization must be performed while satisfying, for all *physical futures contract* $j \in F$, a delivery of L_j kWh. It must also follow the day-ahead market rules. The resulting mathematical expression of the stochastic programming problem is

$$\underset{p,q,f}{\text{maximize}} \sum_{i \in I} \sum_{s \in S} P^s \left[\lambda^s p_i^s - \left(c_i^b + c_i^l p_i^s + c_i^q (p_i^s)^2\right)\right] \tag{17}$$

subject to

$$\sum_{i \in I_j} f_{ij} = L_j \qquad j \in F \tag{18}$$

$$q_i \geq \sum_{j \in F_i} f_{ij} \qquad i \in I \tag{19}$$

$$\underline{P}_i \leq q_i \leq p_i^s \leq \overline{P}_i \qquad i \in I,\ s \in S \tag{20}$$

$$f_{ij} \geq 0 \qquad i \in I,\ j \in F \tag{21}$$

The set of constraint (18) ensures that the energy of the j-th physical futures contract, L_j, will be completely fulfilled among all the committed units. Constraint (19) is the formula that represents the Iberian Electricity Market's rule that forces all of the energy involved in the futures contracts to be bid through the price acceptance bid. The set of constraint (20) expresses the relationships between the price acceptant bid quantity, q_i, the matched energy for scenario s, p_i^s, and the minimum and maximum generation levels $\underline{P}_i$ and $\overline{P}_i$, respectively. Finally, the nonnegativity of the variables f_{ij} is expressed in (21).

The set of formulas (17)–(21) has the form of a concave quadratic maximization problem that can be solved easily by available optimization software; in this case, AMPL (Fourer & Amemiya, 2003) and CPLEX (CPLEX, 2008) were used.

4.2 Results

The results of the optimization process are shown in this section. From the point of view of the GenCo, the main result of the model is the optimal bid curve that has to be sent to the market operator.

Figure 6 shows the optimal bid curves for each committed thermal unit belonging to the GenCo at hour 20, built from the optimal value of the decision variables (Corchero & Heredia, 2011). The bidding curves are optimal in the sense that if the GenCo systematically submits these bids to the day-ahead market, then the expected value of the profits will correspond to the optimal value of (17), and they will maximize expected profits in the long run. As we can see from Figure 6, the optimal bid curves are discontinuous piecewise-linear functions. The first

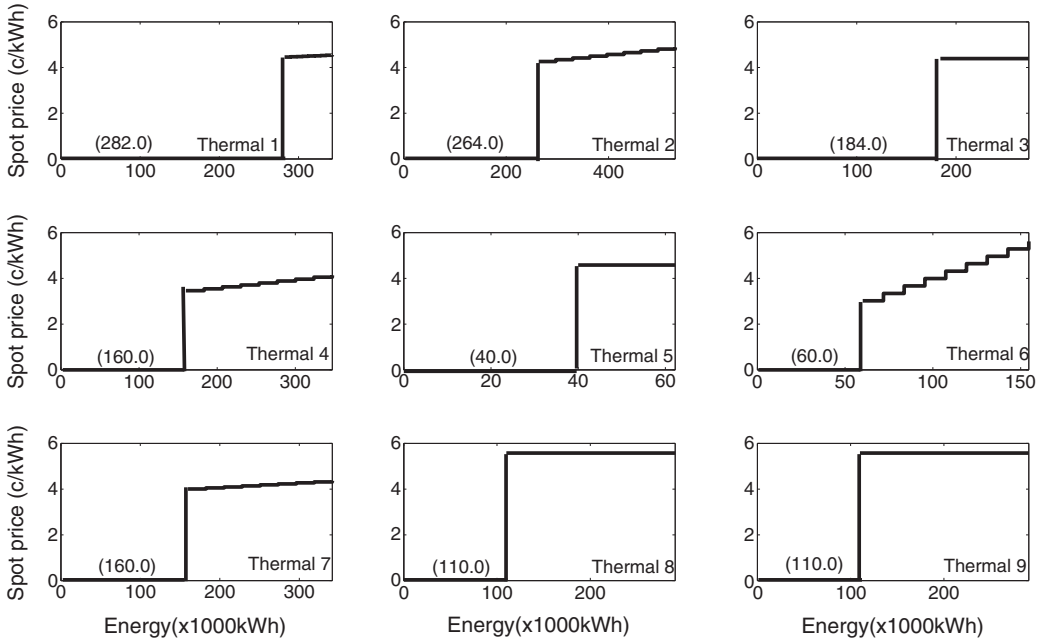

Figure 6. *Bidding curve for each unit in euro cents per kWh.*

interval of the bid curve corresponds to the price acceptant sale bids, that is, the bid with zero price (optimal value of variable q_i). This price acceptance sale bid is indicated in parenthesis as (price, energy). As we have introduced, if a unit has to be committed in the market, it must submit its minimum technical capacity at a zero price so that the clearing algorithm ensures that these units are committed. Notice that there are some thermal units that have bid at a zero price a quantity greater than the minimum power capacity. For instance, thermal unit 7 is offering 160 000 kW for free to the day-ahead market, although its minimum generation amounts to 100 000 kW. This is a direct consequence of the units' participation in the physical futures contract being covered.

5 Conclusions

The main objective of this article was to develop a methodology for developing the optimal bid to be sent to the market operator for each thermal unit at each one of the 24 day-ahead market auctions. This methodology has been designed for the Iberian Electricity Market spot prices, and it is based on a combination of factor analysis and stochastic programming algorithms.

Forecasting electricity prices for the next day is not an easy task because of the complexity of the formation of the actual electricity clearing prices. Electricity prices depend on macroeconomic indicators such as crude oil prices, exchange rates (Muñoz & Dickey, 2009), scarcity of raw materials involved in electricity generation, and internal or external political reasons associated with the stability of a raw material-producing country. Because of this, it is necessary to have a procedure to accurately forecast the electricity prices for the next day and provide it as input to the optimization procedure.

Our approach in this work has been to decompose the electricity prices into a multivariate time series of 24 series, one for each hour of the day, to apply TFSA methodology to reduce the dimensionality of the multivariate time series, and to obtain the factors or latent variables that govern the dynamics of the electricity prices. The TSFA software package has proven to be a very useful tool for achieving this goal.

The forecast procedure based on the factor models gives reasonable forecasts, ones that are equivalent to those obtained from an ARIMA model (Corchero, 2010). However, the advantage of the factor models presented here lies in its simplicity. Building an ARIMA model for the electricity prices requires reliable knowledge of the underlying time-series structure, which is not the case for the procedure presented here. This advantage makes it easy to implement the models automatically so that companies can use it habitually.

Stochastic programming is a powerful methodology that can be used to extend classical optimization problems. In this application, a set of scenarios was used as input to the stochastic programming model to develop the optimal bid to be sent to the market operator. The bid that maximizes the expectation of a GenCo's day-ahead market benefits.

Acknowledgements

The authors would like to acknowledge the many valuable comments and suggestions made by Dr. Carol Blumberg and the Co-Editor-in-Chief, Prof. Vijay Nair, as well as the contributions made by two anonymous referees. Thanks to all of them, the quality of this manuscript has been greatly improved. This work was supported by the Ministry of Science and Technology of Spain through MICINN Project DPI2008-02153. The work of C. Corchero was supported by the Ministry of Science and Technology of Spain through FPI Grant BES-2006-12311.

References

Alonso, A.M., Garcia-Martos, C., Rodriguez, J. & Sanchez, M.J. (2008). *Seasonal dynamic factor analysis and bootstrap inference: application to electricity market forecasting*, Working Paper 08-14. Statistics and Econometric Series 06, Universidad Carlos III de Madrid, Spain.

Anderson, T.W. & Amemiya, Y. (1988). The asymptotic distribution of estimators in factor analysis under general conditions. *Ann. Stat.*, **16**(2), 759–771.

Bai, J. (2003). Inferential theory for factors models of large dimensions. *Econometrica*, **71**(1), 135–171.

Bai, J. & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, **70**(1), 191–221.

Bai, J. & Ng, S. (2008). Large dimensional factor analysis. *Found. Trends. Econ.*, **3**(2), 89–163.

Baillo, A., Cerisola, S., Fernandez-Lopez, J. M. & Bellido, R. (2006). Strategic bidding in electricity spot markets under uncertainty: a roadmap. *IEEE Power Engineering Society General Meeting*. DOI: 10.1109/PES.2006.1708895.

Birge, J.R. & Louveaux, F. (1997). *Introduction to Stochastic Programming.* New York: Springer-Verlag.

Boivin, J. & Ng, S. (2005). Understanding and comparing factor-based forecasts. *Int. J. Cent. Bank*, **1**(3), 117–151.

Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. (1984). *Classification and Regression Trees.* Monterrey, California: Wadsworth Publishing.

Browman, A. & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis: the Kernel Approach with S-Plus Illustrations*. New York: Oxford University Press.

Cattell, R.B., Cattell, A.K.S. & Rhymer, R.M. (1947). P-technique demonstrated in determining psychophysiological source traits in a normal individual. *Psychometrika*, **12**, 267–288.

Cattell, R.B. (1963). The structuring of change by P-technique and incremental R-technique. In *Problems In Measuring Change*, Ed. C. W. Harris, pp. 167–198. Madison: The University of Wisconsin Press.

Chen, X., He, Y., Song, Y.H., Nakanishi, Y., Nakahishi, C., Takahashi, S. & Sekine, Y. (2004). Study of impacts of physical contracts and financial contracts on bidding strategies of GenCos. *Electr. Power Energy Syst.*, **26**, 715–723.

Conejo, A.J., Contreras, J., Espinola, R. & Plazas, M.A. (2005). Forecasting electricity prices for a day-ahead pool-based electric energy market. *Int. J. Forecasting*, **21**, 435–462.

Conejo, A.J., Garcia-Bertrand, R., Carrion, M., Caballero, A. & Andres, A. de. (2008). Optimal involvement in futures markets of a power producer. *IEEE Trans. Power Syst.*, **23**(2), 703–711.

Contreras, J., Espinola, R., Nogales, J. & Conejo, A.J. (2003). ARIMA models to predict next-day electricity prices. *IEEE Trans. Power Syst.*, **18**(3), 1014–1020.

Corchero, C. (2010). *Short-term bidding strategies for a generation company in the Iberian Electricity Market*, PhD Thesis, Universitat Politecnica Catalunya, Spain.

Corchero, C. & Heredia, F.J. (2011). A stochastic programming model for the thermal optimal day-ahead bid problem with physical futures contracts. *Comput. Oper. Res.*, **38**(11), 1501–1512.

CPLEX. (2008). CPLEX Optimization subroutine library guide and reference. Version 11.0. *CPLEX Division, ILOG Inc.,* Incline Village, NV, USA.

Dempster, A.P., Laird, N.M. & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Ser. B.*, **39**(1), 1–38.

Doz, C., Giannone, D. & Reichlin, L. (2012). A quasi maximum likelihood approach for large approximate dynamic factor model. *The Review of Economics and Statistics*, **94**(4), 1014–1024.

Dupacova, J., Consigli, G. & Wallace, S.W. (2000). Scenarios for multistage stochastic programs. *Ann. Oper. Res.*, **100**(1-4), 25–53.

Efron, B. & Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* London: Chapman & Hall/CRC.

Engle, R.F. & Watson, M.W. (1981). A one-factor multivariate time series model of metropolitan wage rates. *J. Amer. Statist. Assoc.*, **76**, 774–781.

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C. & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychol. Methods*, **4**(3), 272–299.

Forni, M., Hallin, M., Lippi, M. & Reichlin, L. (2005). The generalized dynamic-factor model: one-sided estimation and forecasting. *J. Amer. Statist. Assoc.*, **100**, 830–840.

Fourer, R., Amemiya, Y., Gay, D. M. & Kernighan, B. W. (2003). *AMPL: A Modeling Language for Mathematical Programming,* 2nd ed. Pacific Grove, CA: Brooks/Cole-Thomson Learning.

Garcia, R.C., Contreras, J., van, Akkeren & Garcia, J.B.C. (2005). A GARCH forecasting model to predict day-ahead electricity prices. *IEEE Trans. Power Syst.*, **20**(2), 867–874.

Garcia-Martos, C., Rodriguez, J. & Sanchez, M.J. (2007). Mixed models for short-run forecasting of electricity prices: application for the Spanish market. *IEEE Trans. Power Syst.*, **22**(2), 544–552.

Geweke, J. (1977). The dynamic factor analysis of economic time series. Latent variables in socio-economic models. In *Latent Variables in Socio-Economic Models*, Eds. D. Aigner & A. Goldberger, pp. 365–383. Amsterdam: North-Holland.

Gilbert, P.D. & Meijer, E. (2005). Time series factor analysis with an application to measuring money. Research Report No. 05F10, University of Groningen. SOM Research School.

Growe-Kuska, N., Heitsch, H. & Romisch, W. (2003). Scenario reduction and scenario tree construction for power management problems. In *Proceedings of IEEE Bologna Power Tech Conference*, Vol. 3, Bologna, Italy.

Guan, X., Wu, J., Gao, F. & Sun, G. (2008). Optimization-based generation asset allocation for forward and spot markets. *IEEE Trans. Power Syst.*, **23**(4), 1796–1807.

de Haan, J., Leertouwer, E., Meijer, E. & Wansbeek, T. (2003). Measuring central bank independence: a latent variable approach. *Scot. J. Polit. Econ.*, **50**(3), 326–341.

Härdle, W. (1990). *Applied Nonparametric Regression.* Cambridge: Cambridge University Press.

Hastie, T. & Tisbshirani R. (1990). *Generalized Additive Models.* London: Chapman and Hall.

Jungbacker, B., Koopman, S.J. & van der Wel, M. (2011). Maximum likelihood estimation for dynamic factor with missing data. *J. Econ. Dyn. CONTROL*, **35**, 1358–1368.

Karakatsani, N.V. & Bunn, D.W. (2008). Forecasting electricity prices: the impact of fundamentals and time-varying coefficients. *Int. J. Forecast.*, **24**, 764–785.

Kaut, M & Wallace, S.W. (2003). Evaluation of scenario-generation methods for stochastic programming. Technical Report 14, SPEPS Working Paper. DOI: 10.1.1.15.6773.

Koopman, S.J., Ooms, M. & Carnero, M.A. (2007). Periodic seasonal reg-ARFIMA-GARCH models for daily electricity spot prices. *J. Amer. Statist. Assoc.*, **102**(477), 16–27.

Marcelino, M. & Schumacher, C. (2010). Factor MIDAS for nowcasting and forecasting with ragged-edge data: a model comparison for German GDP. *Oxford B. Econ. Stat.*, **72**(4), 305–9049.

Mendes, E.F., Oxley, L. & Reale, M. (2008). Some new approaches to forecasting the price of electricity: a study of Californian market. In *MODSIM 2007 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand,* Eds. L. Oxley & D. Kulasiri. Available at http://ir.canterbury.ac.nz/handle/10092/2069.

McCallum, B.T. (1970). Artificial orthogonalization in regression analysis. *Rev. Econ. Stat.*, **52**(1), 110–113.

Molenaar, P.C.M. (1985). A dynamic factor model for the analysis of multivariate time series. *Psychometrika*, **50**(2), 181–202.

Muñoz, M.P. & Bunk, D.W. (2007). Covariates of stochastic volatility in electricity prices. In *International Symposium on Business and Industrial Statistics,* pp. 141-142. ISBIS-2007*,* University of Azores.

Muñoz, M.P. & Dickey, D.A. (2009). Are electricity prices affected by the US dollar to Euro exchange rate? The Spanish case. *Energ. Econ.*, **31**(6), 857–866.

Nocedal, J. & Wright, S. (2000). *Numerical Optimization.* Springer Series in Operations Research and Financial Engineering. New York: Springer-Verlag.

Pan, J. & Yao, Q. (2008). Modelling multiple time series via common factors. *Biometrika*, **95**(2), 365–379.

Peña, D. & Box, G.E.P. (1987). Identifying a simplifying structure in time series. *J. Amer. Statist. Assoc.*, **82**, 836–843.

Peña, D. & Poncela, P. (2004). Forecasting with nonstationary dynamic factor models. *J. Econometrics*, **119**, 291–321.

Peña, D. & Poncela, P. (2006). Nonstationary dynamic factor analysis. *J. Statist. Plann. Inference*, **136**, 1237–1257.

Proietti, M. (2011). Estimation of common factors under cross-sectional and temporal aggregation constrains. *Int. Statist. Rev.*, **79**(3), 455–476.

Ruszczynski, A. & Shapiro, A. (2003). *Stochastic Programming.* Handbooks in Operations Research and Management Science, vol. 10. Amsterdam: Elsevier.

Sargent, T. & Sims, C. (1977). Business cycle modeling without pretending to have too much a priori economic theory. In *New Methods in Business Cycle Research: Proceedings from a Conference*, Federal Reserve Bank of Minneapolis, pp. 45–109.

Shahidehpour, M., Yamin, H. & Li, Z. (2002). *Market Operations in Electric Power Systems: Forecasting, Scheduling, and Risk Management.* Hoboken, New Jersey: IEEE-Wiley-Interscience.

Schumacher, C. (2007). Forecasting German GDP using alternative factor models based on large datasets. *J. Forecasting*, **26**, 271–302.

Shumway, R.H. & Stoffer, D.S. (1982). An approach to time series smoothing and forecasting using the EM algorithm. *J. Time Ser. Anal.*, **3**, 253–264.

Spanos, A. (1984). Liquidity as a latent variable An application of the MIMIC model. *Oxford B. Econ. Stat.*, **46**, 125–143.

Stevenson, M. (2001). Filtering and Forecasting Spot Electricity Prices in the Increasingly Deregulated Australian Electricity Market. In *International Institute of Forecasters Conference*, Atlanta, June 2001. Available at http://ideas.repec.org/p/uts/rpaper/63.html.

Stock, J. & Watson, M.W. (2002). Forecasting using principal components from a large number of predictors. *J. Amer. Statist. Assoc.*, **97**(460), 1167–1179.

Tsay, R.S. (2010). *Analysis of Financial Time Series*, 3rd ed. Hoboken, New Jersey: Wiley.

Wallace, S.W. & Fleten, S.E. (2003). Stochastic programming models in energy. In *Stochastic Programming*, Eds. A. Ruszczynski & A. Shapiro. Handbooks in Operations Research and Management Science, 10, pp. 637–677. Amsterdam: Elsevier.

Wang, A.J. & Ramsey, B. (1998). A neural network based estimator for electricity spot-pricing with particular reference to weekend and public holidays. *Neurocomputing*, **23**, 47–57.

Wansbeek, T. & Meijer, E. (2000). *Measurement Error and Latent Variables in Econometrics.* Advanced Textbooks in Economics, Vol. 37. Amsterdam: North-Holland.

Watson, M.W. & Engle, R.F. (1983). Alternative algorithms for estimating dynamic factor, MIMIC and varying coefficient regression models. *J. Econometrics*, **23**, 385–400.

Zhu J. (2009). *Optimization of Power System Operation.* IEEE Press Series on Power Engineering. Hoboken, New Jersey: John Wiley and Sons.

Résumé

Dans les marchés libéralisés de l'électricité, la façon dont les producteurs d'électricité gérent leur production est base sur une offre horaire qui est transmise vers le marché day-ahead. Malgré l'objectif des producteurs d'électricité est de maximiser leurs profits, le spot price auquel l'énergie sera payée n'est pas connu au cours du processus d'appel d'offres. Cette situation rend la prévision du spot price un point essentiel dans les stratégies de gestion économique de l'électricité. Dans ce travail, nous appliquons des modéles de prévision des facteurs pour l'estimation du spot price en étudiant sa pertinence. Les résultats obtenus en termes de qualité de l'ajustement sont équivalents á ceux obtenus avec les modéles ARIMA, mais leur application est plus facile. Ce modŁle de prévision du spot price est utilisé pour générer les scénarios des prix d'un modŁle de programmation stochastique qui trouve l'offre optimale de génération dans le marché Espagnol de l'électricité day-ahead Les courbes d'offres optimales de génération sont présentées et analysées.

[*Received January 2011, accepted January 2013*]