Applied Energy 239 (2019) 610–625

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

# Prediction and explanation of the formation of the Spanish day-ahead electricity price through machine learning regression☆

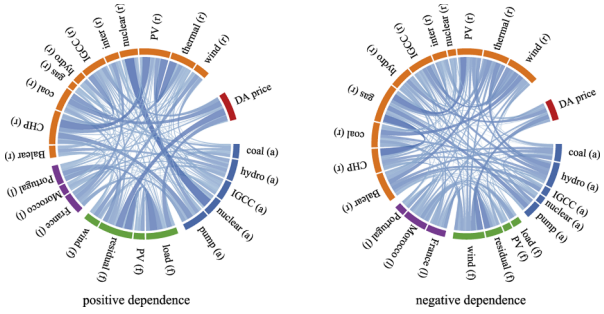
Guzmán Díaz*, José Coto, Javier Gómez-Aleixandre

*Dep. of Electrical Engineering, University of Oviedo, Campus de Viesques, s/n, 33204, Spain*

## HIGHLIGHTS

- We propose a regression-tree-based method for modeling electricity price formation.
- The explanatory variables are extracted from publicly accessible energy related data.
- The energy-related data are free and published by the TSO in a graphical interface.
- The model shows good accuracy in predicting the price formation.
- It also allows for a non-linear analysis of the dependence of price on predictors.

## GRAPHICAL ABSTRACT




## ARTICLE INFO

*Keywords:*
Linear regression
Principal components
Quantile regression
Gradient boosting regression
Day-ahead electricity price

## ABSTRACT

Until recently, detailed information on the power system state to estimate future spot prices by regression analysis was generally restricted to qualified parties. However, to ensure transparency in operation, the Spanish Transmission System Operator has launched an informative web in which a sizable amount of real-time energy-related data can be consulted through a graphical interface. Undoubtedly, this provides the opportunity for non-qualified parties to develop applications and algorithms in which price forecast and maybe knowledge about how price is determined are required.

This paper approaches the use of data extracted from that interface with two aims: the prediction of the day-ahead price in a simple way, and the exploration of the influence that the underlying energy drivers have on it. For the prediction we specified a quantile regression model based on Gradient Boosted Regression Trees. It improves the accuracy over multiple linear regression models at the cost of more complexity, and still it has simpler specification and tuning compared to other machine learning approaches. The calculated metrics show that our model produces remarkably low prediction errors when using the median as point prediction method (RMSE = 2.78 €/MWh, MAE = 1.94 €/MWh, and MAPE = 0.059). Interestingly, the quantile regression model also allows to inherently define prediction intervals, with a different interpretation of accuracy. Our results show that on average 90% of times the prediction error will not exceed 6.8 €/MWh.

We also implemented a partial dependence analysis on that model. This implementation—as far as we know the first time employed to analyze the formation of electricity prices—has shown to be of significant usefulness in detecting highly non-linear relationships.

☆ This work was supported in part by the Spanish Ministry of Economics, Industry, and Competitivity under Grant ENE2016-80053-R.
* Corresponding author.
*E-mail addresses:* guzman@uniovi.es (G. Díaz), jcoto@uniovi.es (J. Coto).

https://doi.org/10.1016/j.apenergy.2019.01.213
Received 7 September 2018; Received in revised form 9 January 2019; Accepted 24 January 2019
Available online 06 February 2019
0306-2619/ 

## 1. Introduction

### 1.1. Motivation

When an increasingly large number and variety of parties are bound to actively participate in the management of energy, ahead knowledge of the electricity price formation and structure becomes a strategic asset. An increasing number of applications by purely price-taker parties demand this knowledge prior to scheduling generation, demand, and storage [1]. For instance, price prediction is key to make the best use of fast-response residential demand, improving the sheer direct control following real-time price signals [2]. Profitability obtained from energy arbitrage, in which storage is used as a means of selling energy previously purchased at inferior prices, is obviously dependent on the price differential between purchasing and generation periods [3]. Some recent algorithmic proposals are highly reliant on the availability of accurate predictions; such as for instance [4], where Chen et al. proposed scheduling microgrid units through a portfolio optimization method, in which eventually the uncertainty focused on the predicted price—and not on other sources of uncertainty such as power output and load demand. In the algorithm presented in [5], price prediction was the fundamental driver to define the aggregation of building loads in a smart grid framework. On the whole, price-sensitive smart grid approaches, following for instance the *cyber-physical* paradigm described in [6], may benefit from price models that allow dynamically predicting prices, and not only from real-time price signals provided by smart meters.

Up until recent years, however, energy generation and demand commitments were fundamentally ascribed to qualified parties with direct participation in the electricity markets; and detailed information about forecasts, system reliability, technical restrictions, or real-time infeeds and demands were by and large accessible only to those major parties. In Spain, that information is received by qualified parties through individual XML messages from the Transmission System Operator (TSO). This situation has historically limited the development of prediction algorithms—other than those based on autoregressive models—by those parties that did not qualify for access to the electricity market.

Therefore, those market parties that do not actively participate in the market bidding may inevitably resort to specialized sources to obtain accurate predictions.

Objectives of transparency have recently changed the landscape and made it easier for relevant information to be now freely accessible to any interested user. This has been the case in Spain, where the TSO has launched a web-based platform (e·sios) where information about the power system past, current, and predicted states is published through a graphical interface. Importantly, it can be consulted freely; see for instance [7].

This is also the case of, for instance, Elia in Belgium [8], Fingrid in Finland [9], RTE in France [10], EirGrid in Ireland [11], SVK in Sweden [12], or APG in Austria [13].

The idea motivating this paper is that web data extraction along with a necessarily simple but sufficiently accurate regression model can be used to predict the electricity price by subjects that do not qualify to receive individual, informative messages from the TSO. Also, data available so far in e·sios are not the whole set received by qualified parties in real-time, so the idea in this paper is to see how these data suffice to build an accurate model.

### 1.2. Literature review

There is a large body of literature on electricity price characterization. A thorough revision is provided by Weron in [14], who identifies six groups of models. The models discussed in this paper are in the category of Statistical models (Regression models), except that regression trees (and other machine learning methods) were not contemplated in his paper. A complementary review for this purpose can be found in [15].

The review in [14] provides a thorough analysis of methods for forecasting electricity day-ahead prices. That work was published in 2014. In what follows, we focus on more recent publications related to our approach.

The simplest and most widely studied models are the linear regression models. One of the two models employed in [16] was a multivariate time series regression model using historical data. A variation of this theme appears in [17], where the authors used individual autoregressive models of the explanatory variables, which were integrated into one single model through a linear regression (indeed a lasso estimation, which improves accuracy of the linear regression through variable selection). This allowed the authors to obtain probabilistic predictions through Monte Carlo simulations of the explanatory variables. In both cases they applied a well-known technique, broadly used in econometry. But by its nature it was limited to linear relationships, related only to the mean of the price, and sensitive to outliers and dependence among variables. Some of these problems can be mitigated through dedicated techniques. For instance, when multicollinearity is detected by means of variance inflation factors or condition indices, alternative models based on regularized regressions can be considered. Or the skewness introduced by outliers can be reduced by means of robust regression. Or non-normality can be possibly tackled by transforming one or more variables or resorting to non-linear regression models. But these transformations require expert judgment and do not always give the expected results (in this respect, see the influential paper by Dormann et al., where a detailed analysis is performed [18]).

In principle, machine learning techniques such as artificial neural networks, support vector machines, or random forests, to cite some, are less sensitive to the problems shown by traditional linear regression techniques. For instance, random forests inherently emulate regularization through the number of variables sampled at each split. But in general, it is the different approach that makes the result different. Whereas in linear regression the focus is on determining the impact of some variables on others—reflected in the coefficients of the regression—machine learning methods focus on how some variables can be used to predict others. This different view theoretically improves the prediction power, but at the cost of losing explanatory power.

As an example of artificial neural networks (ANN), Singh et al. characterized New South Wales' market price using a sort of recurrent ANN [15]. Also along these lines, Chen et al. employed a recurrent ANN, although their focus was on solving a portfolio optimization problem [4]. In both cases, one of the major shortcomings, summarized in [15], was the complexity of training. As a response to this problem, the recurrent ANN was also approached in [19] but with a variant Long Short Term Memory (LSTM) employed to simplify the ANN learning. In these studies, the ANN regressed the price on previous occurrences, and external predictors were not employed. Panapakidis et al. proposed six different artificial neural networks models, some of them including exogenous variables, to make point predictions of the day-ahead price with errors lower than those obtained using time-series models [20]. This higher precision was confirmed in the meta-analysis performed in [21], where their methods based on machine learning always outperformed those based on statistical methods (mostly autoregressive processes) except in one particular case (functional autoregressive with exogenous variables models, fARX).

A methodology combining ANNs with a number of other approaches was presented in [22]. ANN, wavelet analysis, and an underlying) autoregressive moving average (ARMA) model were employed together, resulting in an accurate but appreciably complex model. Complexity, as a visible feature because of the number of demanded techniques to

achieve a proper model tuning, can also be perceived in [23], where again as in [22] wavelets were combined with ANNs. In [23] a heuristic optimization algorithm was introduced as a means of improving an ANN's learning process.

Although the best machine learning technique cannot be chosen a priori, possibly a simpler alternative to ANNs may be the regression trees. In [24] a comparison between both approaches was tested to predict energy demand in buildings. The approach based on regression trees demonstrated to be only marginally less accurate than ANNs. However, regression trees are generally simpler to calibrate and faster once trained. In the field of price prediction, Gaillard et al. tested a regression-tree-based model in the context of the Global Energy Forecasting Competition [25]. The model was ranked first in the competition, hence showing the potential of these models.

Finally, a number of studies have recently examined the use of probability forecasting—meaning the use of prediction intervals and probability densities rather than a mean value of the prediction—as a way of additionally informing about the error incurred in the predictions. Moreover, one of the advantages of potentially having a forecast distribution functions is that simulations can be conducted to generate scenarios [26,27]. A useful review can be found in [28], where it was emphasized that only recently, from 2016 onwards, have probabilistic methods been gaining popularity compared to the prevalent favored point prediction [14]. The reason for using probability predictions can be found in [29], where Nowotarski and Weron justified that it is not possible to select the best price forecast model based only on point predictions. Thus, in [29] point predictions from several models were employed to produce prediction intervals.

Quantile regression was first introduced by Bunn et al. in [30] to investigate electricity prices, demonstrating its value compared to GARCH (generalized autoregressive conditional heteroskedasticity) and CAViAR (Conditional Value at Risk by Quantile Regression) methods. It is also reported in [31], where the Nord Pool's electricity price was regressed on temperature, wind speed, precipitation, and irradiance. The authors claimed that by using a quantile regression, their analysis outperformed others that had been based on the mean value. In [32], the authors demonstrated that the results from a quantile regression model could be effectively used to test the price sensitivity through scenario analysis.

### 1.3. Aims and contributions

This paper demonstrates that regression models built from web accessible data sources containing the state evolution and forecast of a power system can provide reasonably accurate price predictions to interested parties. In this respect, the two main aims of this paper are (i) to discuss the use and interpretation of a gradient boosting regression tree (GBRT) model of the electricity price formation and prediction using those data sources; and (ii) by using that model, analyze the price drivers in the Spanish markets.

More particularly, the aims and contributions of this paper are summarized as follows:

- This paper proposes using publicly-accessible data to regress the day-ahead price on a number of electrical power predictors. We employed data not from a dedicated repository but from a graphic-oriented web page; where trends, in-feeds, and predictions are represented as charts by the Spanish system operator in real-time. Although a specific API must be defined to obtain those data in numeric format, they have the advantage that they are very rich, meaning that a huge amount of variables can be accessed at the same time.
- This paper also shows the implementation of a gradient boosting regression tree (GBRT) model to characterize the price formation, combining many simple models to produce an accurate ensemble model [33]. Regression trees were already employed in [25] to predict electricity prices. However, the model in [25] was fairly complex, because of the underlaying submodels employed as predictors. Although we alternatively included more predictors, because of the variety offered by e·sios, our purpose was to balance complexity and accuracy, so that this approach could be replicated by price takers searching fairly accurate predictions in non-bidding approaches.
  Price prediction models may be classified into two broad categories. The first category would include those models that look backward in time to infer rules that drive the prices: from linear autoregression through feed-forward neural networks (see for instance the comparison in [34] to get an overall picture). The technique we have employed lies in a second category. Although we have adapted autoregressive features, the model is mainly based on explaining the price using exogenous regressors (see for instance [35], for a concept explanation). Price evolution evident or hidden features are not the basis. Instead, it is the combination of exogenous variables that gives the price. Importantly, this offers the advantage of interpreting the price formation from the power system state.
  In this sense, we also regressed the price on the available exogenous regressors by means of a principal component regression (PCR) model. This type of models avoid the multicollinearity problems associated to the conventional multiple linear regression (MLR) models, at the cost of interpretability. In our study, where high multicollinearity was expected in advance because of the large number of higly correlated predictors employed, the PCR model serves as a "best" attempt to regress the price by conventional methods. The GBRT model is more complex to implement than the PCR. Still, compared to other proposed techniques [22,23], the GBRT approach shows to be simpler and appears to be sufficiently accurate for a price taker needs. Its results, on the other hand, have shown to be far superior to the MLR and PCR models.
- From the point of view of risk assessment, prediction intervals seem more adequate. We therefore have implemented quantile-based GBRTs, rather than mean-based regression, to analyze the uncertainty in the price formation. Particularly, we provide evidence about the particular asymmetry observed in the prediction intervals, highlighting the difference to methods that yield symmetrical intervals based on the point prediction as the center of an interval.
- Finally, this paper analyzes the dependence of price on the individual predictors. Unlike MLR models, machine learning methods (black-box models) do not have a direct equivalent to the analysis of coefficients to assess the price effects brought on by changes in the predictors. We alternatively introduced partial dependence as a way of visualizing this sensitivity.
  To our knowledge, this non-linear analysis of the sensitivity of electricity prices to exogenous variables is not found in the literature so far.
  Although more laborious to obtain than the simpler liner regression coefficients, it has shown to provide clearer conceptual and numerical insights about the price formation than the simpler regression coefficients. The use of partial dependence has unveiled nonlinear and piecewise relations, which serve to explain the underlying influence of the energy driver on the price formation.

### 1.4. Structure of the study

The remainder of the paper is structured as follows. First, in Section

2, a description of the employed regression models is presented. Also, a description of the web extracted data is provided. In Section 3, the results are reported, with some explanations about the model performance metrics, and the use of variable importances and partial dependences as measures of the underlying influence of the predictors. In Section 4, a discussion about the results is presented. Finally, Section 5 concludes the paper and summarizes it.

## 2. Data and methods

### 2.1. Regression models

#### 2.1.1. Multiple linear regression (MLR) and principal component regression (PCR)

Regression analyses offer an effective way of relating the response of a dependent variable to a set of observations. The most direct way to conduct such analyses is by means of a multiple linear regression (MLR). In those analyses, the model employed is a linear one of the form

$$y = \beta_0 1 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon = \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta} + \varepsilon; \tag{1}$$

where $y$ is the dependent variable, the vector $\mathbf{x}_i$ encapsulates the dependent variables, $\varepsilon$ is the disturbance term, and $\boldsymbol{\beta}$ are the regression coefficients to be determined at the stage of calibrating the model using $N$ pair of response and explanatory observations, $\{y_k, \mathbf{x}_k\}_{k=1,\ldots,K}$. In our case, $y_t$ is the day-ahead price of the $t$-th hour (thus $T = 24$ hours) and $\mathbf{x}$ are the $p$ available explanatory variables—the power forecasts, availabilities, and in-feeds freely accessible from e·sios. Parameters in $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ can be simply obtained from solving

$$\min_{\boldsymbol{\beta}\in \mathrm{IR}^p} \sum_{k=1}^{K} (y_k - \mathbf{x}_k^{\mathsf{T}}\boldsymbol{\beta})^2. \tag{2}$$

There are certain problems with the use of multiple linear regression, however. One of these is that with highly correlated variables the coefficient estimation becomes unstable, since the diagonal of the Cholesky factor coming from the correlation matrix degenerates to almost zero. This produces instability at the time of inversion of the factor. As we show later, the publicly available data related to energy contains variables that are inherently highly correlated. Consequently, using multiple linear regression for predicting prices using the available data is problematic.

The effect of correlation between variables, resulting in the multicollinearity problems, can be dealt with through several methods. For instance, regularized regression (including least absolute shrinkage and selection operator, LASSO, and ridge regression) allows subsetting the variables so that the prediction accuracy is improved. Another way, which we have employed in this paper to overcome the multicollinearity problem is to use principal component analysis (PCA), which avoids the inversion of the matrix causing the parameter estimate instability. The transformation can be achieved by means of a singular value decomposition, projecting the original variables $\mathbf{x}$ onto a new orthogonal basis, where each component represents the maximum variability of the original data in decreasing order of magnitude. The projected vectors, the *component scores*, are thus a linear combination of the original variables weighted by the *loadings*.

PCA avoids the problem of multicollinearity related to the high correlation of the independent variables, because the components are by definition orthogonal. This allows performing a multiple linear regression of the day-ahead price on the principal components, on the basis that these are but a rotation of the original variables [36]. Unfortunately, it also removes one of the most interesting capabilities of the multiple regression analysis, which is the use of the $\boldsymbol{\beta}$ coefficients to investigate the influence of each independent variable on the day-ahead price. Nonetheless, this is a minor drawback if the aim is to use the model for prediction.

#### 2.1.2. Gradient boosting regression trees

To compare the accuracy as a function of the data, we also implemented a regression tree model. These type of models divide the original data space into smaller subspaces in which the interrelations are easier to model. Though programmatically more complex, the procedure is easy to interpret, which is one of the advantages of the regression trees in modeling complex interrelations.

Regression trees represent an improvement over MLR and PCR approaches when the data structure is complex because of the partitioning of the space into simpler subspaces. The space partitioning is unknown in advance, and must be the subject of an iterative search that determines the optimal structure that combines simplicity and accuracy. Still, their use may be inefficient and lead to overfitting because all terms in the tree are assumed to interact. Variables are not considered to act independently, even though they have weak interactions. We therefore employed a further refinement using gradient boosting regression trees, as an ensemble of regression trees that allow weak interactions to be modeled.

Supervised learning methods aim at producing a mapping $f: \mathscr{X} \mapsto \mathscr{Y}$, where $\mathscr{X} \subset \mathrm{IR}^p$ is the predictor space and $\mathscr{Y} \subset \mathrm{IR}$ is the response space. The objective is to estimate a mapping $\hat{f}(\mathbf{x})$, such that a loss function $\mathscr{L}(y, f(\mathbf{x}))$ evaluated using known paired observations $(\mathbf{x}_k, y_k)$, $k = 1, \ldots, K$, is minimized:

$$\hat{f}(\mathbf{x}) = \underset{f(\mathbf{x})}{\operatorname{argmin}} \sum_{i=1}^{N} \mathscr{L}\Big(y_i, f(\mathbf{x}_i)\Big) \tag{3}$$

In the case of GBRT, the model is additive using simpler decision trees as the weak models or learners. The model is built starting from a proposal $f_m(\mathbf{x})$, in a sequence $m = 1, \ldots, M$, that predicts $y$. In a subsequent step of the sequence, the model is improved by adding an estimator $h_m(\mathbf{x})$ that corrects the previous tree:

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + h_m(\mathbf{x}), \tag{4}$$

where $h_m(\mathbf{x})$ is the weak learner; in this case a regression tree. Because this subsequent tree must improve the previous tree estimate, the procedure at step $m$ yields a fitting of $h_m(\mathbf{x})$ to the residual $y - f_m(\mathbf{x})$. That is, at each step the algorithm improves the prediction by narrowing the residual difference.

The GBRT considers an additive expansion trained in this forward stage-wise manner to approximate the response of the original set by means of the basis functions $h_m(\mathbf{x})$. However, in practice finding the solution to this problem may be difficult. That is why usually $h_m(\mathbf{x})$ is chosen to be the most parallel to the negative gradient along the observed data. To that purpose, rather than the observed data a set of calculated *pseudo-residuals* is employed following this gradient line.

The final model is therefore a weighted sum of the $M$ intermediate trees

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^{M} \gamma_m h_m(\mathbf{x}), \tag{5}$$

where the weight is estimated by minimizing the loss function $\mathscr{L}(y, f(\mathbf{x}))$. This loss function measures the difference between the prediction $f_m(\mathbf{x})$ and the observations $y$. Several formulations can be decided, as for instance the classic squared error, being this a matter of simple trial and error [37].

**Algorithm 1.** Outline of the GBRT technique

**Input** : response $y$ and explanatory $\mathbf{x}$ variables
**Output** : optimal tree
**Init** : $f_0(\mathbf{x}_m) = \text{argmin}_\beta \sum_{k=1}^{K} \mathcal{L}(y_i, \beta)$;
**for** $m \in \{1, \ldots, M\}$ **do**
  **for** $k \in \{1, \ldots, K\}$ **do**
    **Define** : $r_{km} \leftarrow \left[-\frac{\partial \mathcal{L}(y_k, \mathbf{x}_k)}{\partial f(\mathbf{x}_k)}\right]_{f(\mathbf{x})=f_{m-1}(\mathbf{x})}$, the pseudo-residuals;
  **end**
  **Fit** : a weak learner $h_m(\mathbf{x})$ to $\{(r_{km}, \mathbf{x}_k)\}, \quad k = 1, \ldots, K$;
  **Calculate** : $\gamma_m \leftarrow \text{argmin}_\gamma \sum_{i=1}^{N} \mathcal{L}(y_k, f_{m-1}(\mathbf{x}_k) + \gamma\, h_m(\mathbf{x}_k))$;
  **Update** : $f_m(\mathbf{x}) \leftarrow f_{m-1}(\mathbf{x}) + \gamma_m\, h_m(\mathbf{x})$
**end**
**return** $\hat{f}(\mathbf{x}) \leftarrow f_M(\mathbf{x})$

This GBRT-based fitting procedure is outlined in Algorithm 1.

### 2.2. Data model

The Spanish TSO is Red Eléctrica de Espaàa (REE). To achieve transparency of information goals, REE has set up an information system named Sistema de Información del Operador del Sistema (System Operator Information System), e·sios, in which historical as well as real time and forecast data are presented [38]. Access to these data can be done in two ways, either through a repository containing historical data [39] or through a web page where time evolution of variables of interest can be interactively selected [40]. In both cases, data are freely accessible. Data in the repository [39] can be accessed in Extensible Markup Language (xml) format bundled into zipped files named "hemerotecas". By contrast, data in [40] describe real-time data through a graphical interface. There are no delays in the available data in this case, and some of the data are indeed real-time events. We employed this data source.

A drawback with these datasets is that they are meant to be represented online as line plots (see for instance [7]), a format useful to analyze trends and make comparisons between different variables. However, APIs can be developed to query e.sios url to get JSON data (see [41]). This way, we downloaded them in tabular form after requesting a free token from REE. Particularly, we employed a short Python code to access these data.

We selected four categories of explanatory variables, based on the idea that these should be the variables that a market agent would be more interested in consulting at the moment of issuing a bid. In fact, it is the composition of the perceptions obtained from key information about the power availability and forecasts that eventually produce the different predictions and aggregate in the market bids to produce a given price. So we employed:

(a) The available non-renewable generation: thermal (nuclear, IGCC, and coal) and hydro. These indeed are the dispatchable aggregated power. Availability does not entail ensured dispatch, as their schedule will depend on their merit order and technical restrictions. Therefore, this category helps the market agent to make an estimate of the so-called "thermal gap."
(b) The available forecasts at a global level. Although in Spain the renewable-energy generators must have their own forecast to produce a correct bid that avoids later penalization in the unbalance market, REE issues an aggregated forecast of PV and wind energy for the 24-h period of the next day. This forecast is supplemented by a load forecast in a finer time scale. Both forecasts combine to produce the residual forecast.
(c) International link capacity. This is an estimate of the power exchange with the four neighboring countries (France, Portugal, and Morocco). Depending on its sign it may be treated as generation (import) or consumption (export).
(d) The above three categories allude to an estimate of power availability and exchanges. Differently, the last category we considered informs about the evolution of the power generation over the past hours. It is kind of a "memory" of the trajectory of the generation system.

As it can be seen, these variables lie in the space of decision of a market agent. A decision is based on how the power generation has evolved and how it is going to do in the foreseeable future. This precisely is the idea of the regression model: To emulate how the system evolved in the past when conditions at the time of bidding where similar to what they are now.

Both analyzed models, PCR and GBRT, rely on the analysis of a sample of $K$ pairs $(y_k, \mathbf{x}_k)$. It is of crucial importance to realize that in e·sios dataset the variables are recorded as a function of the time at which the event occurred or for the time it was forecast. This means that these data must be referred to a common time, which we selected to be 11.00 p.m., the gate closure hour for submitting bids for the $D + 1$ day.

We split the vector of explanatory variables into four vectors, each corresponding to the four categories explained above. The $N$ pairs are then split into the tuple $(y_k, \mathbf{x}_{ak}, \mathbf{x}_{fk}, \mathbf{x}_{\ell k}, \mathbf{x}_{rk})$. Vectors informing about the available dispatchable generation and international links, $\mathbf{x}_a$ and $\mathbf{x}_l$, can be categorized as real-time observations. However, the power forecasts and trajectories, $\mathbf{x}_f$ and $\mathbf{x}_r$, refer to future and past observations, respectively.

The ensemble prediction of the 24 day-ahead prices of day $D + 1$ must be done before gate closure at 11.00 a.m. on day $D$. However, past observations, or historical data, are stored in e·sios reflecting the time for which they were issued. This means that load, wind, and PV forecasts for hour $t$ on day $D + 1$ are stored under that time stamp, although they were published at 11.00 p.m. on day $D$. This must be taken into account at the time of specifying the regression model, in the same way that observations *before* 11.00 p.m. must be brought to the common time, which is 11.00 p.m. on day $D$.

With this in mind, we specified a regression model in which the day-ahead price for hour $t$ can be related to the publicly available data as:

$$y = \boldsymbol{\beta}_\ell^{\mathsf{T}} \frac{\mathbf{x}_\ell}{\mathrm{B}^{t+13}} + \boldsymbol{\beta}_{\mathrm{f}}^{\mathsf{T}} \frac{\mathbf{x}_{\mathrm{f}}}{\mathrm{B}^{t+13}} + \boldsymbol{\beta}_{\mathrm{a}}^{\mathsf{T}} \mathbf{x}_{\mathrm{a}} + \left( \sum_{i=0}^{s} \boldsymbol{\beta}_{\mathrm{r},i}^{\mathsf{T}} \mathrm{B}^i \right) \mathbf{x}_{\mathrm{r}} + \beta_t t, \tag{6}$$

where the back-shift or lag operator B implies that

$$\mathrm{B}^i x_t = x_{t-i}. \tag{7}$$

In (6), $\mathbf{x}_{\mathrm{f}}$ is the vector of forecast variables (first column of Table 1), $\mathbf{x}_\ell$ is the vector of international links (first column of Table 1), $\mathbf{x}_{\mathrm{a}}$ is the vector of unit availability variables (second column of Table 1), and $\mathbf{x}_{\mathrm{r}}$ is the vector of real-time generation and load variables (third column of Table 1). The first three terms are common for any of the 24 predictions of the next day. They account for the certain knowledge about the system status and trajectory that an agent has at 11.00 p.m. Particularly, the third term makes use of the back-shift operator to define an

**Table 1**
Variables making up the model.

| Forecasts | Availabilities | Real-time |
|---|---|---|
| Day-ahead price | Coal power | Total load |
| Wind power | Hydro power | Coal power |
| PV power | IGCC power | Hydro power |
| International links | Nuclear power | IGCC power |
| Total load | Pumping hydro power | Nuclear power |
| Residual load | | Fuel power |
| | | Wind power |
| | | PV power |
| | | Solar thermal power |
| | | International links |
| | | CHP |
| | | Balear link |

**Table 2**
Original table, as obtained from e·sios web page.

| Day | Hour | Price | Forecasts | Availabilities | Real-time |
|---|---|---|---|---|---|
| 1 | 0:00 | $p_0$ | $\mathbf{f}_0$ | $\mathbf{a}_0$ | $\mathbf{r}_0$ |
| … | … | … | … | … | … |
| 1 | 9:00 | $p_9$ | $\mathbf{f}_9$ | $\mathbf{a}_9$ | $\mathbf{r}_9$ |
| 1 | 10:00 | $p_{10}$ | $\mathbf{f}_{10}$ | $\mathbf{a}_{10}$ | $\mathbf{r}_{10}$ |
| 1 | 11:00 | $p_{11}$ | $\mathbf{f}_{11}$ | $\mathbf{a}_{11}$ | $\mathbf{r}_{11}$ |
| … | … | … | … | … | … |
| 2 | 0:00 | $p_{24}$ | $\mathbf{f}_{24}$ | $\mathbf{a}_{24}$ | $\mathbf{r}_{24}$ |
| 2 | 1:00 | $p_{25}$ | $\mathbf{f}_{25}$ | $\mathbf{a}_{25}$ | $\mathbf{r}_{25}$ |
| … | … | … | … | … | … |

autoregression of the trajectory of the system until 11.00 p.m. The last two terms, by contrast, modify the prevision as a function of the time for which it is made in the future. It is in the second-to-last term where the observations about forecasts stored in e·sios' dataset are translated using the back-shift operator to provide a consistent calibration. Finally, the last term introduces a dummy regression of the time for which the prediction is made.

### 2.3. Practical implementation

We wrote the above model in tabular form, with each column representing an element of vector $\mathbf{x}$. Each row represented an observation of the type described in (6).

We imported data from e·sios web and classified them as shown in Table 1. These data come with a date tag or label, indicating the hour at which they are valid. Therefore, there is a difference between forecasts and real-time records. For instance, the power supplied by an agent at for 11:00 p.m. on day $D$ is recorded using that date tag. However, the forecast issued by REE at 11:00 p.m. on day $D$ about the power at 8:00 a.m. on day $D + 1$ is recorded with the tag corresponding to this latter hour (8:00 a.m. day $D + 1$ and not 11:00 a.m. day $D$). This is the reason to rearrange all data to produce a single event—a table row—in which the data that was available to the market agents at 11:00 a.m. is shown.

To exemplify the procedure we followed, we show two generic tables. Table 2 refers to the date-tagged data, classified according to Table 1. This is the original data downloaded from e·sios web site. Table 3 shows how the data is finally rearranged. This latter is the table that would later be processed by the GBRT and PCR algorithms.

To obtain Table 3 from Table 2, first, we moved the forecasts (prices, load, renewable, and international links) to the actual hour, and set the hour-ahead tag to the time corresponding to the data location. It must be recalled, the data on the server is stored not with a time label corresponding to the time at which the forecast was made, but by using a tag corresponding to the time for which the forecast was made. Thus, in the headers of Table 3 there are two hour tags. The first corresponds to the actual hour at which the price prediction is made. Therefore, it is always equal to 11:00 a.m. The second tag, "hour ahead" is the code identifying the day-ahead hour for which the prediction was made, and it is the hour that will be later considered in this paper when analyzing the impact of time on the price prediction. Accordingly, $p_{24}$ and $\mathbf{f}_{24}$, which are the realization and forecast of prices for the first hour of day 2 appear as the first records of Table 3. This procedure completes the first two terms of (6) and define the dependent variable, the day-ahead price $y$.

Real-time and available units are "copied" in Table 3 without modification. That is because they represent the information available to the agents at 11:00 a.m. That is the reason why $p_{24}$ appears in Table 3 in the same row as $\mathbf{a}_{11}$: the agent information at 11:00 of day 1 includes the $p_{24}$ forecast for the ahead-hour 0 of day 2 (indicated by means of a zero in the last column, first row of Table 3) and the information about available units at 11:00 a.m. of day 1, $\mathbf{a}_{11}$.

The same procedure is applied to obtain the value of real-time power, $\mathbf{r}_{11}$. But additionally, we proceeded to build an autoregression model by bringing the past observations to the target hour 11:00. This means that the final table has as many new columns as lead values introduced. In the example, there are two past observations: leads 1 and 2. This makes that the new two columns including in the first row $\mathbf{r}_{10}$ and $\mathbf{r}_9$. Also, this accounts for the fourth term in (6).

The second row modifies just the price and forecast values, compared to the first row, and the value of the hour ahead (in the second row it is a 1, indicating that the agent at 11:00 of day 1 had information regarding hour 1 forecasts of day 2). Once the 24 h of day 2 forecasts and prices serve to complete the first 24 rows of Table 3, the procedure is repeated using the information that an agent would have at 11:00 of day 2. The final table is in the form of Table 3, except that the first two columns are removed.

We employed data for the period 1 January 2015 to 31 December 2017. This led to a matrix of 26,269 rows. When we introduced a 3-h lag in the definition of $\mathbf{x}_r$ the matrix of explanatory variables had 66 columns.

We then subset the table by rows, creating three sets: train, validation, and test. The sub-setting of the data was done randomly (no need for sequential records, because the autoregressive information was already contained in the real-time columns), in a proportion 60%, 20%, and 20% (train, validation, and test). We did not conduct cross-validation tests because of the data size.

We conducted the above procedures on the R environment, which we found really suited for the task. Additionally, we employed the open-source machine learning platform H2O [42], whereby we created a 3 Gigabyte server instance to proceed with all computations.

In the case of the GBRT model, we conducted a hyperparameter search [43] to obtain a set of optimal parameters. Particularly, we selected as varying hyperparameters the maximum depth to which each

**Table 3**
Final data arrangement using data in Table 2.

| Day | Hour | Price | Forecasts | Availabilities | Real-time | Real-time lead 1 | Real-time lead 2 | Hour ahead |
|---|---|---|---|---|---|---|---|---|
| 1 | 11:00 | $p_{24}$ | $\mathbf{f}_{24}$ | $\mathbf{a}_{11}$ | $\mathbf{r}_{11}$ | $\mathbf{r}_{10}$ | $\mathbf{r}_9$ | 0 |
| 1 | 11:00 | $p_{25}$ | $\mathbf{f}_{25}$ | $\mathbf{a}_{11}$ | $\mathbf{r}_{11}$ | $\mathbf{r}_{10}$ | $\mathbf{r}_9$ | 1 |
| … | … | … | … | … | … | … | … | … |
| 2 | 11:00 | $p_{48}$ | $\mathbf{f}_{48}$ | $\mathbf{a}_{35}$ | $\mathbf{r}_{35}$ | $\mathbf{r}_{34}$ | $\mathbf{r}_{33}$ | 0 |
| … | … | … | … | … | … | … | … | … |

**Table 4**
Summary of hyperparameters employed over 23 rounds, to optimize the 50-percentile GBRT model. (For space reasons, only the first and last rounds are presented.)

| | Range | 1 | 2 | 3 | 4 | 20 | 21 | 22 | 23 |
|---|---|---|---|---|---|---|---|---|---|
| Max. depth | 10–75 | 71 | 23 | 44 | 61 | 75 | 75 | 30 | 29 |
| Row sample rate | 0.2–1.0 | 0.39 | 0.57 | 0.56 | 0.24 | 1.00 | 1.00 | 0.90 | 1.00 |
| Col. sample rate | 0.2–1.0 | 0.99 | 0.76 | 0.47 | 0.70 | 0.29 | 0.83 | 0.49 | 0.86 |
| Col. sampling rate/tree | 0.2–1.0 | 0.40 | 0.38 | 0.61 | 0.95 | 0.86 | 0.69 | 0.65 | 0.73 |
| Col. sampling rate/level | 0.2–1.8 | 0.34 | 0.51 | 0.35 | 1.40 | 1.80 | 1.80 | 1.80 | 1.80 |
| Min. rows | 1–16 | 13 | 12 | 12 | 8 | 1 | 3 | 3 | 5 |
| No. bins (cont.) | 16–1024 | 943 | 442 | 652 | 432 | 786 | 388 | 305 | 754 |
| No. bins (categ.) | 16–4096 | 2990 | 1091 | 851 | 2804 | 2857 | 1069 | 341 | 1210 |
| RMSE | | 3.86 | 3.65 | 3.72 | 3.88 | 2.93 | 2.87 | 2.94 | 2.91 |

tree would be built, the row and column sample rates, the column sampling rate for each tree, the relative change of the column sampling rate for every level, the minimum number of observations for a leaf to split, and the number of bins to be included in the histogram and then split at the best point. The range of these variables is shown in Table 4.

We additionally set the learning rate equal to 0.05, with an annealing equal to 0.99. Each model was scored after 10 trees, and the stopping metric selected was RMSE, with the stopping criterion being set to a tolerance of $10^{-3}$ €/MWh after two consecutive rounds.

### 2.4. Calibration metrics

To assess the performance of the calibrated model, we calculated several commonly-used metrics, which we briefly define here. MAE indicates the average magnitude of the errors, without considering their direction.

$$\mathrm{MAE} = \frac{\sum_{k=1}^{K} |\hat{y}_k - y_k|}{K}, \tag{8}$$

where $K$ is the number of samples. RMSE, defined as

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{k=1}^{K} (\hat{y}_k - y_k)^2}{K}}, \tag{9}$$

also measures the average magnitude of the error, but as a quadratic scoring rule it "penalizes" the largest deviations. This may seem advantageous at the time of investigating economic losses due to bad price predictions, but in turn RMSE demonstrates to have some serious disadvantages compared with MAE, as it is discussed in [44]. The MAPE is defined as

$$\mathrm{MAPE} = \frac{1}{K} \sum_{k=1}^{K} \left| \frac{\hat{y}_k - y_k}{y_k} \right|. \tag{10}$$

Note that MAPE indicates the relative absolute deviation in per unit values. It can be obtained as a percentage by simply multiplying by 100.

### 2.5. Explanatory analysis

The importance of the explanatory variables in forming the model is defined as

$$I_i = \sqrt{\mathrm{E}_{\mathbf{x}} \left[ \frac{\partial \hat{f}(\mathbf{x})}{\partial x_i} \right]^2 \times \mathrm{Var}[x_i]}, \tag{11}$$

which in the case of piecewise constant approximations derived from a decision tree $T$ must be approximated by the surrogate

$$\hat{I}_i^2(T) = \sum_{t=1}^{J-1} \hat{i}_t^2 \mathbf{1}\left(v_t = 1\right). \tag{12}$$

This means that the predictor importance is determined as a function of how often the variable is selected during the splitting in the tree building process ($v_t$ is the splitting variable associated with non-terminal node $t$), weighted by how much the squared error improved as a result ($\hat{i}_t^2$ is the empirical improvement measure). Over the $M$ trees $T_M$, the variable importance is obtained as:

$$\hat{I}_i^2 = \frac{1}{M} \sum_{m=1}^{M} I_i^2(T_m). \tag{13}$$

In addition to the relative importance of each explanatory variable, it seems interesting to analyze the sensitivity of the day-ahead price to variations of those variables. However, unlike in multilinear regression models, no simple parametric description exists in GBRT models comparable to the examination of MLR coefficients. Using the explanatory vector $\mathbf{x} = x_1, \ldots, x_p$, the GBRT model can estimate the day-ahead price in general form as

$$\hat{y} = \hat{f}(x_1, \ldots, x_p). \tag{14}$$

But being $\mathbf{x}$ a high dimensional vector applied over an algorithmic function, that prediction is uninterpretable. In [45], Friedman proposed using a marginalized surrogate of $\hat{f}(\mathbf{x})$ in a lower dimensional space, which would show the *partial dependence* between the prediction and a selected variable. The idea behind partial dependences is to compute the average

$$\phi(x_j) = \frac{1}{K} \sum_{k=1}^{K} F\left(x_{1k}, \ldots, x_{(j-1)k}, x, x_{(j+1)k}, \ldots, x_{pk}\right), \tag{15}$$

on the basis that $\phi_j(x)$ will show how the variable $x_j$ influences the predictions of the model upon averaging out the influence of all the other variables. In other words, partial dependence does not ignore the influence of other predictors, but integrates those predictors out.

### 2.6. Summary

A summary of the methodology is outlined in Fig. 1. Basically, the procedure is as follows. Data sets are parsed from e·sios' interface. Data are processed from Tables 2 and 3 format. The ensuing data set is split into three subsets. The training set is employed to train the model, using a set of initial hyperparameters. Then the model is scored—using the RMSE as prefered metric—by means of the he validation data set. If budget is available (with budget we mean kind of access to computational resources, once the validation performance so far is assessed) then another round is conducted using a new set of hyperparameters. If not, the model with best validation score is chosen as the final model. This is utilized to further analyze the performance with the reserved test set and to launch the explanatory analysis (partial dependence and variable importance).

## 3. Results

Table 10 summarizes statistical characteristics of the variables employed in the analysis. Day-ahead price is the explanatory variable,

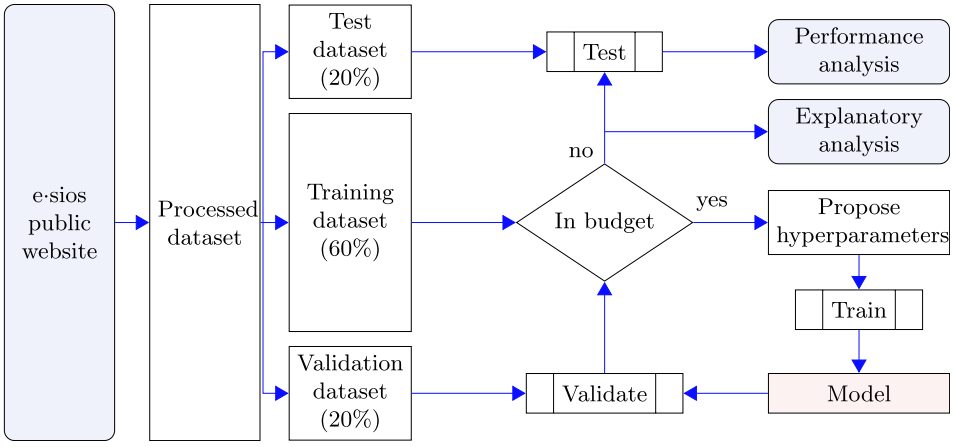



**Fig. 1.** Summary of the methodology.

which is measured in €/MWh. The remaining are explanatory variables, all measured in MWh. Only one instance of each real-time predictor is presented, though in the final model autoregressive variables are incorporated to the model. Nevertheless, the discrepancy observed between the statistical figures of the lags is negligible.

Additional statistical information is provided in Fig. 2, where correlation between all variables is represented using different shades.

Table 4 shows the range of hyperparameters employed during the Bayesian search and a representative set of rounds in order of occurrence. For space reasons, only the first four and last four search rounds are shown. Data corresponds to the of the 50-percentile GBRT model, and the RMSE shown corresponds to the validation set.

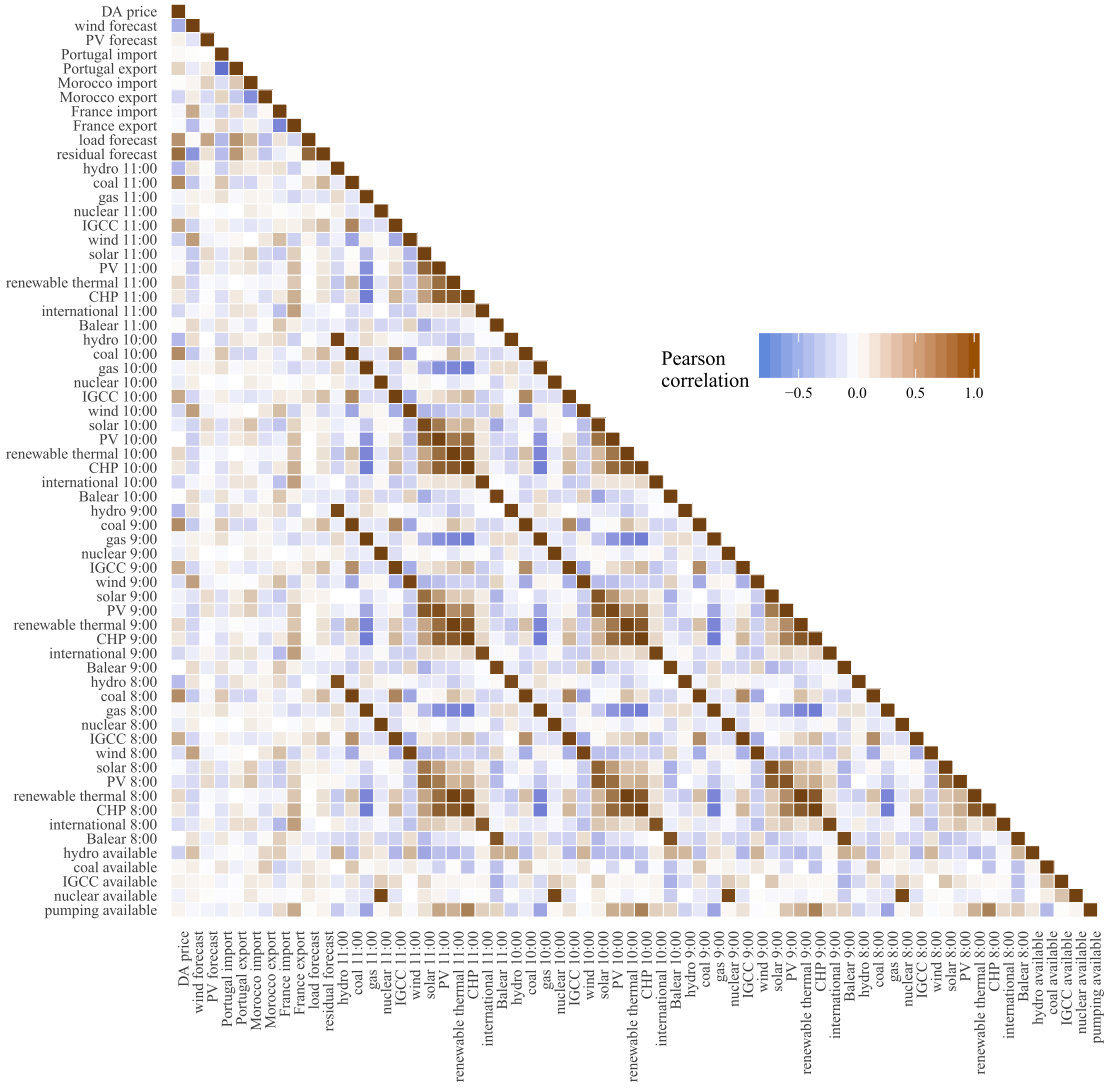



**Fig. 2.** Pearson correlation between the variables employed in the GBRT and PCR models.

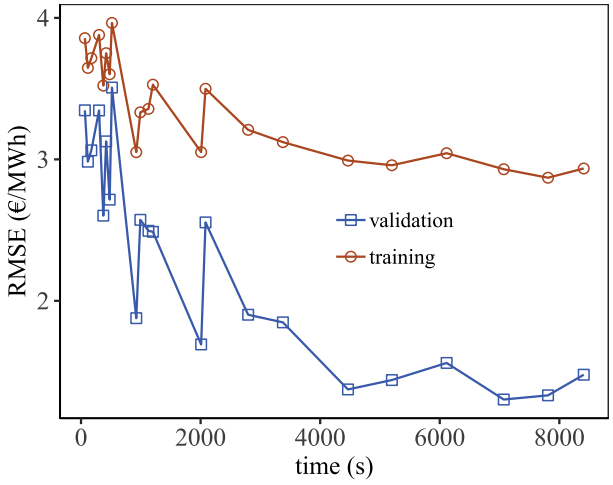



**Fig. 3.** Errors in the training and validation sets over the Bayesian hyperparameter search.

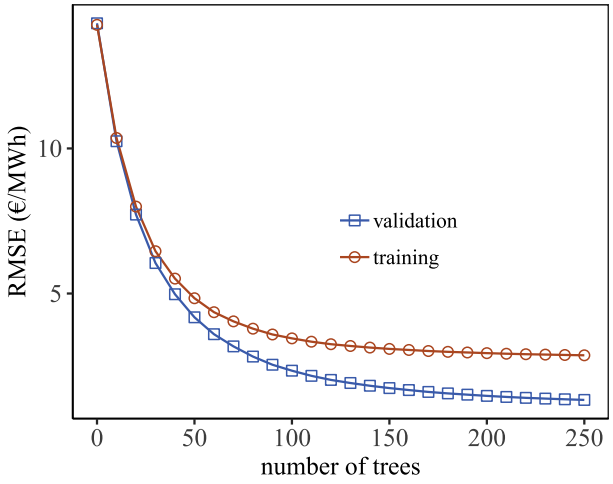



**Fig. 4.** RMSE of the validation and training sets as a function of the number of trees, during the calibration of the 50-percentile best model.

**Table 5**
Performance metrics of PCR model.

| | RMSE | MAE | $R^2$ | MAPE |
|---|---|---|---|---|
| Train | 4.878 | 3.803 | 0.886 | 0.346 |
| Validation | 17.648 | 15.269 | 0.623 | 0.337 |
| Test | 14.226 | 11.045 | 0.560 | 0.352 |

**Table 6**
Performance metrics of the GBRT 50 percentile model.

| | RMSE | MAE | $R^2$ | MAPE |
|---|---|---|---|---|
| Train | 1.332 | 0.495 | 0.991 | 0.022 |
| Validation | 2.870 | 1.971 | 0.959 | 0.063 |
| Test | 2.783 | 1.943 | 0.964 | 0.059 |

**Table 7**
Performance metrics of the GBRT 95 percentile model.

| | RMSE | MAE | $R^2$ | MAPE |
|---|---|---|---|---|
| Train | 9.380 | 6.577 | 0.570 | 0.288 |
| Validation | 9.446 | 6.753 | 0.560 | 0.282 |
| Test | 9.627 | 6.881 | 0.568 | 0.295 |

**Table 8**
Performance metrics of the GBRT 5 percentile model.

| | RMSE | MAE | $R^2$ | MAPE |
|---|---|---|---|---|
| Train | 9.426 | 6.952 | 0.565 | 0.130 |
| Validation | 9.599 | 7.254 | 0.546 | 0.142 |
| Test | 9.685 | 7.209 | 0.563 | 0.139 |

Fig. 3 serves to compare the evolution of the RMSE in both the validation and the training sets over the hyperparameter search of the 50-percentile GBRT model. Particularly, Fig. 4 focus on the case of the best model—found in round 22 of the hyperparameter search—featuring the process of calibration of a particular model for an increasing number of trees.

The same data are repeated for the 95 and 5 percentile in Tables 8 and 7.

Tables 5–9 show the root-mean-square error (RMSE) and mean absolute error (MAE) value in €/MWh.

Table 9 concentrates on the test set and shows the prediction errors evaluated at each of the 24 h.

Using only the test set, Fig. 5 provides a view of the dispersion in the predictions, as well as a comparison between the 50-percentile GBRT and the PCR model accuracies. A perfect prediction would follow the diagonal line. In Fig. 6, the same representation is used to assess the accuracy—on the training set again—between three GBRT models, calibrated to provide predictions based on the 10, 50, and 90 percentiles.

The relative influences of the individual predictors, $x_i$, on the variation of the approximation $\hat{f}(\mathbf{x})$ are analyzed in Fig. 8. This figure shows a selection of the predictors with largest importance. The value of $I_i$ is represented for the 20 variables that have a contribution to the total model higher than 12%. They are additionally classified into the four major groups described above.

The results of calculating partial dependences are plotted in Figs. 9–11. In the case of Fig. 9, the result is discrete because the explanatory variables involved, week day and hour, are categorical. In Fig. 10 the four panels serve to analyze the four families of publicly available data proposed above (forecasts, link availability, dispatchable generation, and recent power share). Finally, Fig. 11 illustrates the case of recent power share in more detail, describing the partial dependence between price and the hours preceding the forecast at 11.00 a.m.

## 4. Discussion

Fig. 2 visually represents the dependence among the main drivers of the day-ahead price, split into the four data categories considered in this paper (plus the day-ahead price). A visibly major correlation is observed between the availability and operation of must-run nuclear units. Not so between available and employed coal units, which shows a weak relationship, due to the cost-driven operation of these units. Some relationships between different technologies are also evident. Thus, real-time PV and thermal (both solar) are obviously highly correlated. Forecasts of total and residual demand are also notably correlated. Other dependencies, however, may be thought as spurious—or at least of difficult explanation. That is the case of the high correlation between solar PV and CHP; when this latter in Spain is mostly sold by means of bilateral contracts. Finally, some dependencies do show interesting relationships, which though of difficult direct explanation provide ground for developing an expert prediction system. See for instance, the negative high relationship between CHP and gas units, or the positive correlation between the demand forecast and the Portugal-Spain exchange link.

The relationship between price and other variables is also readily observed in those charts. The price is highly positively correlated with the residual demand forecast, the running coal units and, to a lesser extent, the IGCC units. On the other hand, it is negatively correlated

**Table 9**
Accuracy of PCR and 50-percentile GBRT models as a function of the hour for which the prediction is made.

| | Hour | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RMSE | PCR | 8.31 | 8.09 | 8.06 | 8.11 | 8.24 | 10.35 | 12.58 | 13.63 | 13.58 | 13.10 | 12.68 | 12.48 |
| | GBRT | 1.74 | 1.59 | 1.62 | 1.71 | 1.65 | 1.87 | 1.98 | 1.57 | 1.68 | 1.51 | 1.38 | 1.10 |
| MAE | PCR | 6.17 | 5.84 | 5.67 | 5.63 | 5.90 | 8.03 | 9.94 | 11.15 | 11.12 | 10.71 | 10.53 | 10.49 |
| | GBRT | 0.96 | 0.84 | 0.75 | 0.71 | 0.89 | 1.01 | 1.07 | 0.89 | 0.95 | 0.80 | 0.74 | 0.66 |
| | Hour | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 |
| RMSE | PCR | 11.91 | 11.14 | 10.99 | 11.13 | 12.94 | 15.32 | 16.12 | 15.49 | 14.12 | 12.15 | 10.81 | 9.39 |
| | GBRT | 1.05 | 1.03 | 1.25 | 1.18 | 1.64 | 2.15 | 2.10 | 1.95 | 1.83 | 1.90 | 2.05 | 2.34 |
| MAE | PCR | 9.92 | 9.00 | 8.77 | 8.82 | 9.99 | 11.81 | 12.99 | 12.98 | 11.71 | 9.73 | 8.86 | 6.93 |
| | GBRT | 0.58 | 0.56 | 0.69 | 0.66 | 0.96 | 1.25 | 1.15 | 1.15 | 1.00 | 1.15 | 1.38 | 1.34 |

with wind forecast and real-time hydro.

### 4.1. Model performance

Table 4 shows a summary of how hyperparameters affected the accuracy of the model. Its performance is assessed by means of the RMSE metric. The evolution of this metric can be observed in Fig. 3 for the training and validation sets. In this table, however, only the validation metric is represented as this was the selection criterion to decide which was the best model. (Generally, all models demonstrated higher errors in the validation set scoring than in the training set calibration. See Fig. 4.)

The combined analysis of Figs. 3 and 4, and Table 4 provides insight on how the calibration was conducted and the performance trend in the model. Fig. 3 represents the hyperparameter search over 23 rounds. Each round consisted of a model calibration using a set of hyperparameters selected by the search algorithm. A summary of those hyperparameters over the different rounds can be observed in Table 4. In turn, each of the model calibrations consisted of a search through an increasing number of trees (see Fig. 4), until a satisfactory value of the RMSE was found. We selected the RMSE of validation set as the stopping criterion, stopping the search when the variation of this value was below a preset tolerance.

The trend in the RMSE value is perceived as decreasing in Fig. 3 both for the training and the validation sets. It is not smooth, however, demonstrating the peakedness of the loss function. The evaluation of each possible set of hyperparameters was costly, and depended on the complexity of the underlying model—particularly on the number of trees required to achieve a reduced RMSE variation. That cost can be observed in the horizontal axis of Fig. 3, where the accumulated calibration time is represented.

Data in Table 4 reveal the complexity of the loss function. Particularly not many trends can be observed about which the best range of hyperparameters are. For instance, a maximum depth equal to 71 gives a high validation RMSE (3.86), but a tree of depth equal to 75 produced the best fitting. Only the row sampling rate and the column sampling rate per level demonstrated a clear consistency, with higher values being preferable.

In Tables 5 and 6 the RMSE was employed as stopping criterion. In general, the PCR performed much worse than the 50-percentile GBRT. Worse, a large difference can be observed in the RMSE of the three employed sets. This leads us to conclude that the PCR model was overfitted (see for instance the relatively good score of the training set compared to the greater value of the unknown test set). This overfitting

**Table 10**
Summary of variables employed.

| Variable | Minimum | 25-percentile | Mean | Median | 75-percentile | Maximum | Stdev | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|
| DA price | 2.30 | 39.64 | 47.40 | 48.06 | 56.61 | 101.99 | 14.36 | −0.30 | 0.65 |
| Wind forecast | 237.00 | 2952.00 | 5430.36 | 4825.00 | 7295.00 | 17430.00 | 3173.98 | 0.82 | 0.17 |
| PV forecast | 0.00 | 0.00 | 847.96 | 94.20 | 1739.30 | 3650.20 | 1077.29 | 0.89 | −0.73 |
| Portugal import | 0.00 | 1593.00 | 2520.71 | 2574.20 | 3471.90 | 6070.00 | 1299.85 | −0.07 | −0.64 |
| Portugal export | 0.00 | 1323.80 | 2230.65 | 2210.10 | 3104.40 | 6090.40 | 1193.22 | 0.10 | −0.67 |
| Morocco import | 0.00 | 940.00 | 1176.90 | 1320.00 | 1400.00 | 1515.30 | 293.42 | −0.94 | 0.07 |
| Morocco export | −299.70 | 60.00 | 232.61 | 100.00 | 360.00 | 1200.00 | 239.19 | 1.26 | 0.56 |
| France import | −750.00 | 0.00 | 1044.87 | 0.00 | 1706.00 | 6400.00 | 1652.52 | 1.56 | 1.22 |
| France export | −300.00 | 1600.00 | 2843.87 | 2700.00 | 4500.00 | 6400.00 | 1799.54 | −0.08 | −1.17 |
| Load forecast | 17599.00 | 24692.00 | 28591.54 | 28812.00 | 32051.00 | 41514.00 | 4568.29 | 0.07 | −0.89 |
| Residual forecast | 6587.20 | 18206.90 | 21706.24 | 21584.60 | 25354.10 | 37606.00 | 5019.69 | −0.00 | −0.37 |
| Hydro real time | −1455.17 | 2460.50 | 4138.00 | 3943.00 | 5547.17 | 10877.50 | 2286.86 | 0.47 | −0.17 |
| Coal real time | 658.17 | 3173.00 | 5196.57 | 5732.33 | 7070.83 | 9157.00 | 2339.21 | −0.32 | −1.15 |
| Gas real time | 0.00 | 0.00 | 28.97 | 0.00 | 0.00 | 472.00 | 99.73 | 3.23 | 8.75 |
| Nuclear real time | 3740.83 | 5908.67 | 6325.38 | 6728.83 | 7069.67 | 7112.17 | 824.49 | −0.79 | −0.59 |
| IGCC real time | 456.67 | 1929.50 | 3641.13 | 2859.83 | 4554.33 | 15158.67 | 2400.31 | 1.38 | 1.65 |
| Wind real time | 250.50 | 2104.33 | 4751.34 | 3960.67 | 6634.83 | 15865.83 | 3338.76 | 0.99 | 0.41 |
| Solar thermal real time | 0.00 | 66.83 | 925.68 | 724.50 | 1817.83 | 2213.50 | 827.77 | 0.27 | −1.55 |
| PV real time | 0.00 | 1347.50 | 1875.18 | 2254.83 | 2557.67 | 3305.00 | 927.63 | −0.99 | −0.29 |
| Renew. thermal real time | 0.00 | 373.33 | 390.74 | 430.67 | 464.33 | 638.33 | 174.75 | −1.28 | 0.81 |
| CHP real time | 0.00 | 3010.83 | 2913.24 | 3372.33 | 3592.50 | 4014.67 | 1211.16 | −1.84 | 1.75 |
| International real time | −4517.50 | −368.83 | 748.51 | 780.83 | 1861.83 | 4615.17 | 1637.97 | −0.20 | −0.21 |
| Balear real time | −304.00 | −230.00 | −172.13 | −175.00 | −115.17 | 0.00 | 70.30 | −0.17 | −0.77 |
| Hydro available | 10924.60 | 12557.10 | 13134.10 | 13253.50 | 13825.50 | 14486.90 | 800.58 | −0.49 | −0.76 |
| Coal available | 5693.70 | 8089.20 | 8730.48 | 8818.70 | 9419.10 | 10402.80 | 862.60 | −0.36 | −0.32 |
| IGCC available | 18651.70 | 21777.00 | 22626.42 | 22965.20 | 23479.40 | 24559.60 | 1290.97 | −0.75 | 0.04 |
| Nuclear available | 4004.60 | 6105.90 | 6479.62 | 7117.20 | 7117.20 | 7117.20 | 825.18 | −0.94 | −0.45 |
| Pumping available | 1174.80 | 2265.00 | 2679.50 | 2813.90 | 3091.30 | 3364.00 | 474.78 | −0.61 | −0.68 |

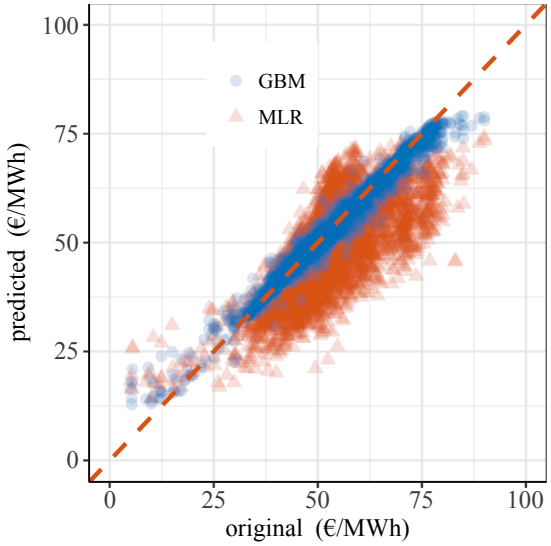



**Fig. 5.** Comparison between original and predicted data sets using a gradient boosting model with $\alpha_{50}$ and a multiple linear regression based on principal components of the original data.

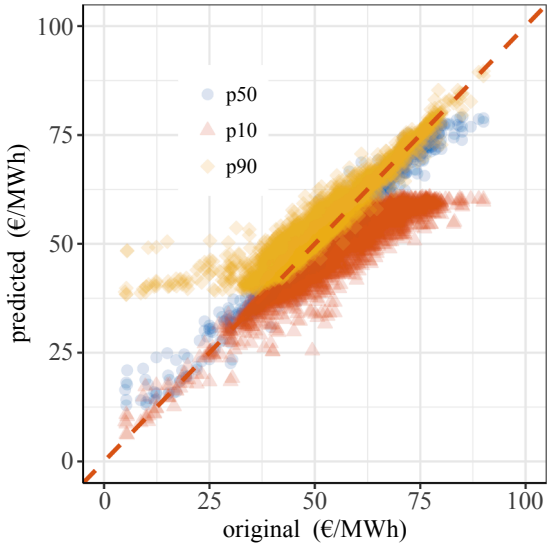



**Fig. 6.** Prediction over the same data set as in Fig. 5. Percentiles $\alpha_{10}$, $\alpha_{90}$, and (again) $\alpha_{50}$ are shown.

also affected the 50-percentile GBRT model. Differences between training and the other two sets are not so remarkable. But his could be attributed to the dependence of RMSE on the sample size (training set used 60% of the total data, with the remaining 40% being equally split between validation and test). In this case, the low value of training MAE compared to the validation value reveals the problem of over-fitting. However, validation and test sets performed similarly in all metrics, despite their two different roles (whereas the former was employed to score the model during the calibration, the latter was unknown to the model at the stage of calibration). This suggests that the model still has good predictive power, despite a degree of possible over-fitting.

Regarding the RMSE, an additional relative measure of performance consists in the comparison of the RMSE with that of the mean value of the dependence variable. This is usually considered a naif prediction. In the case of the validation set, the RMSE of $\bar{y}_k = 47.4$ €/MWh was 14.25 €/MWh. This indicates that the accuracy of the PCR model was even worse than that of the simpler mean prediction: $\frac{17.65}{14.25} = 1.24$. However, the 50-percentile model showed a relative RMSE equal to $\frac{2.87}{14.25} = 0.20$ that of the mean prediction.

The values of $R^2$, the coefficient of determination, serve to show the percentage of the dependent variable variation that is explained by the model. Because the GBRT model is not linear, the employ of this metric may be subject to criticism. But still, it serves to provide an approximation. Again, the PCR model is visibly worse than the 50-percentile GBRT model. Particularly, the 50-percentile GBRT model shows to explain about 96% of the variability of the model.

Tables 8 and 7 refer to the GBRT model, but particularly to extreme quantiles. Compared to the 50-percentile model (Table 6), the accuracy of those models seem at first remarkably worse. However, this is due to the nature of the models. The metrics in the tables refer to point predictions, which is not the object of the extreme quantiles. What it shows is that the median prediction performs well as a point prediction. In the case of the 5- and 95-quantiles, the metrics can be used to represent the mean deviation from the central prediction. On average, it shows that the confidence intervals are symmetric. Not so (as later shown in Fig. 5) at every hour. But on average, those predictions have a mean error about 6.8 €/MWh. We interpret this result as a 90% certainty that the price prediction absolute error will be lower than that figure.

A last measure offered is the MAPE. This widely-used, unitless metric produces a relative value of the error in the model. It is, however, of reduced use in the case of quantile regression, since this metric tends to favor systematically those models whose predictions are low [46]. This is demonstrated by comparing Tables 7 and 8. Although other metrics indicate the similarity in performance of both models, the MAPE indicates a lower error of the 5-percentile model, because its predictions are naturally lower.

To put it into context, the above metrics can be compared to others published in the literature. Although a fine comparison is difficult because, as stated by Weron [14], there is not an 'industry standard' and the error benchmarks used in the literature are often different, a general view can be obtained on the good performance of the model presented in this paper. We particularly refer here to the 50-percentile GBRT model as a point prediction method.

For instance, in [21] an out-of-sample MAPE about 0.13 was reported using deep learning techniques. Szkuta et al. reported in [47] RMSE between 0.64 and 2.00 $/MWh using artificial neural networks, for seven daily predictions. Also employing neural networks to process EPEX prices along with selected external factors, the average RMSE presented in [48] was 6.58 €/MWh. The out-of-sample metrics using ARMA models on data from Lepitzig spot market in [49] obtained RMSE scores in the interval 3.9–10.0 €/MWh. The minimum MAE was 2.6 €/MWh. In [50], a comparison between several techniques was conducted using EPEX data. The MAE results varied from 3.20 €/MWh for 1-h-ahead forecast through 5.53 €/MWh for 24-h-ahead predictions. In [51], using back-propagation neural networks the MAPE was as low as 0.070 for one-step-ahead and 0.059 for six-step-ahead predictions of prices from Australia and France markets. Using wavelet transforms as the core of their method, the authors of [52] reported minimum MAPE of 0.063 and maximum of 0.119, depending on the selected month, on data from PJM. The lowest MAPE reported in [53] from a combination and comparison of autoregressive methods employing Nord Pool data was 0.07, and the maximum was 1.22.

Particularly about Spanish electricity market, some results are as follows. Using data from 2015 and employing a combination of techniques (wavelet transform, extreme learning, and ARMA), the range of MAPE obtained in [22] was from a maximum 0.129 (January) to a minimum 0.019 (July). The average MAPE in [50] after using feed-forward neural networks was 0.063. Uniejewski et al. tested different variance stabilizing methods on neural networks for different markets [54]. In the case of the Spanish market, the lowest reported MAE was 5.84 €/MWh. Finally, in [34] a proposed feed-forward neural network was compared to 17 other published methods. Except for ARIMA, mixed-model, and neural network models, they were hybrids of either

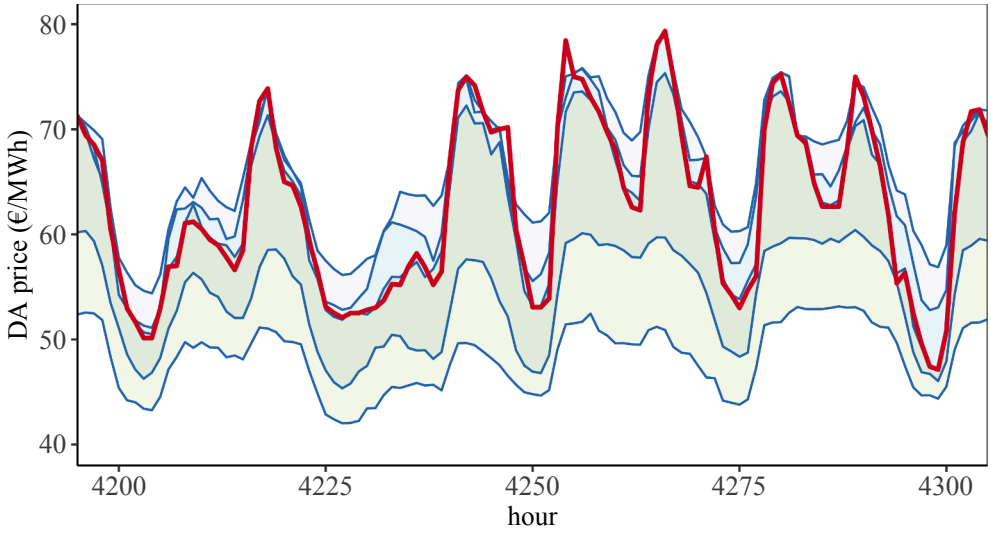



**Fig. 7.** Quantile predictions of the day-ahead price in June 2017. Quantile predictions are shown with different shades in four bands (they are best observed at $h = 4275$). The band on top correspond to the prediction between 95 (upper bound) and 90 (lower bound) percentiles. Next below are (90, 50), (50, 10), and (10, 5) percentiles. The solid, red line is the actual DA price.

hard or soft computing models. The reported MAPE ranged from 0.051 through 0.100.

The above cases are difficult to compare with ours exactly, since the metrics and benchmarks vary. In the case of RMSE and MAE, they are absolute metrics and hence affected by the range of variation of the prices. This is particularly observed in the analysis by Zeil et al. on 10 markets [35], where their LASSO method obtained RMSE ranging from 5.17 (Poland) through 29.94 (Denmark) €/MWh, and MAE from 2.68 (Poland) through 6.78 (Belgium) €/MWh. Moreover, the RMSE is affected by the size of the sample. In the case of MAPE, low prices favor better scores. Still, it is clear that the errors observed in our results (RMSE = 2.78, MAE = 1.94, and MAPE = 0.059) are remarkably low.

A question that was raised during this research was whether the furthest predictions—those made for the last hours of $D + 1$ day—would be lest accurate than those in the first hours. We analyzed it by working out the MAE and RMSE of the test frame (Table 9). Both regression models differentiated the times for which the prediction are calculated by introducing factors indicating the week day and the hour of the day at the time of calibrating the model using the training and validation frames, as shown in (6). Consequently, at the prediction time—using the test frame to assess the model accuracy—we can pinpoint the hour by introducing the code of that precise hour. Using the hour code as a filtering criterion, the RMSE and MAE can be computed.

Additionally, Table 9 reveals that furthest predictions were less accurate when the more accurate 50-percentile GBRT model was employed. The least accurate predictions were those made for 18:00 onwards. This is different in the case of the less accurate PCR model, in which the latest predictions were oddly enough amongst the more accurate. Indeed, the PCR model prediction errors do follow the demand curve, being large at the times of highest demand. Maybe because the 50-percentile GBRT model does not follow a linear relationship with demand, its prediction error shows to be less related to that demand and the model compensates for the large weight that the demand has on the overall prediction.

The above assessment about accuracy can be complemented by observing the scatter plots in Figs. 5 and 6. These are convenient representations in that they provide insight into the symmetry of the predictions, which the RMSE and MAE do not. For instance, Fig. 5

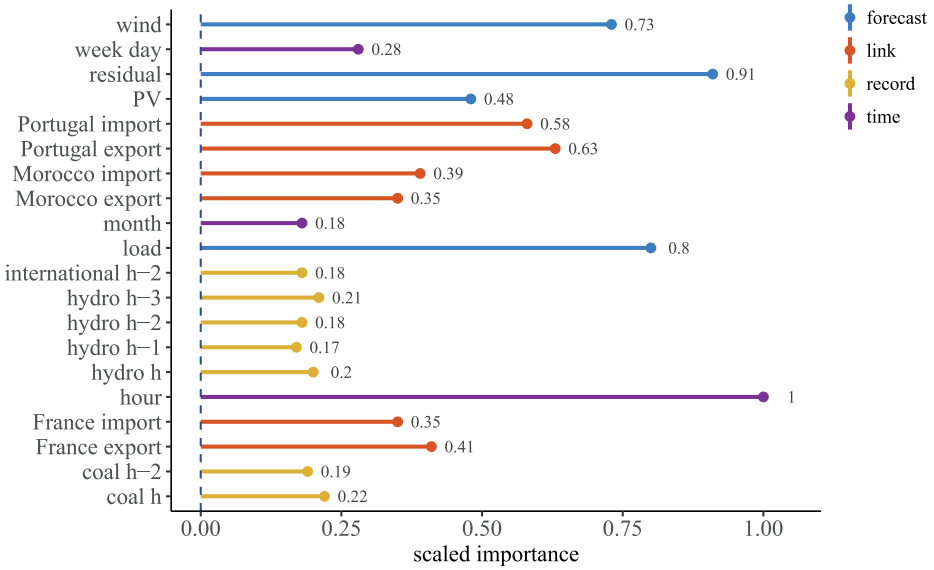



**Fig. 8.** Variable importance (only the most important, 20 out of 66, are represented) of the percentile 50 prediction model.

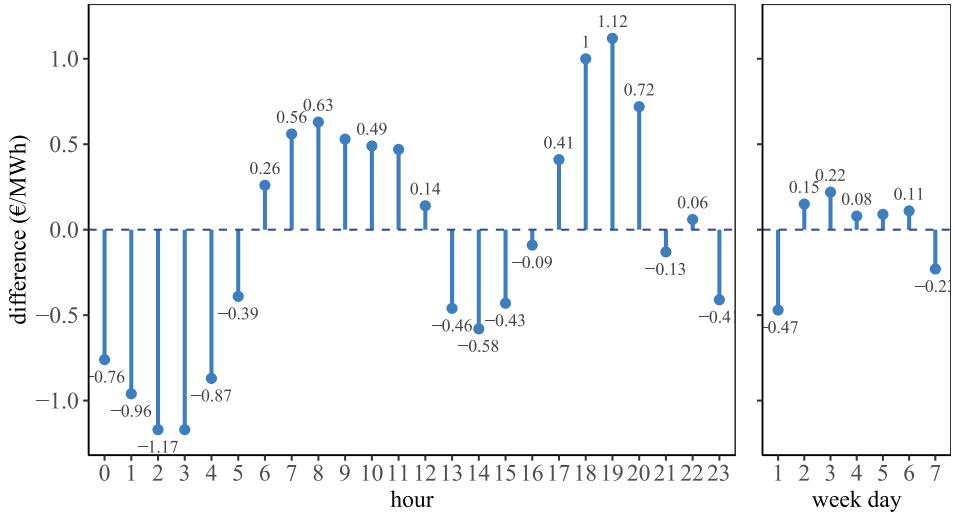



**Fig. 9.** Partial dependence between day-ahead prices and the categorical variables related to the prediction hour and week day in the 50-percentile GBRT. For regularization purposes, the partial dependence shown is indeed *the deviation* of the partial dependence from the mean value of the prediction.

reveals that the worse accuracy of the PCR model was due to short predictions. Consistently, except in the range from 50 through 60 €/MWh, the forecast was mostly below the diagonal line, indicating a conservative prediction. The distribution of forecast errors was thus positively skewed. This was in stark contrast with the performance of the 50-percentile GBRT model, which provided a more balanced prediction with a reduced skewness. Practically in all cases (see for instance the extreme lowest and highest prices) the 50-percentile GBRT model was more accurate than the PCR model.

Although one of the aims of quantile prediction is to provide prediction intervals rather than direct point forecasts, an important result of this research has been that the 50% quantile is an accurate point prediction of the day-ahead price. In the comparison of the differences among quantile predictions provided in Fig. 6, visibly the $\alpha_{50}$-based model was the best GBRT model if a point prediction is desired. In the case of the $\alpha_{90}$, the predictions were visibly high, with particularly high errors in the low-price range. This situation was reversed when $\alpha_{10}$ was employed. In this case, the largest errors occurred in the high-price end. On the whole, $\alpha_{50}$ predictions performed reasonably well in all the price range.

The idea of employing the 50% quantile forecast for point predictions is further analyzed in Fig. 5. Along with the observed price realization (thick, solid line) four confidence intervals based on quantile estimates are represented. The quantiles are $\alpha_{95}$, $\alpha_{90}$, $\alpha_{50}$, $\alpha_{10}$, and $\alpha_5$. So each of the four confidence bands correspond to the four probabilistic forecasts $\text{pr}\{90 \leqslant y \leqslant 95\}$, $\text{pr}\{50 \leqslant y \leqslant 90\}$, $\text{pr}\{10 \leqslant y \leqslant 50\}$, and $\text{pr}\{5 \leqslant y \leqslant 10\}$. The observed price realization (the actual value) falls close to the line defining $\alpha_{50}$. Particularly, this is observed in the way in which the observed (unknown) price realization neatly divides the shaded areas into two below and two above.

It is important to realize that these results are in contrast with the most popular definition of prediction intervals, which in based on point forecasts and their corresponding error [28]. The point forecast is in most cases employed as the center of an interval, the bounds of which are determined by the two quantiles $\frac{\alpha}{2}$ and $\left(1-\frac{\alpha}{2}\right)$ in a symmetric way. In other cases, by adopting a Gaussian assumption about the residuals, as in [55], or bootstrapping its variance [56], the results are again symmetrical confidence intervals. By contrast, in Fig. 7 the confidence intervals are visibly asymmetric. For high prices the confidence interval defined by ($\alpha_{50}$, $\alpha_{90}$) is visibly narrower than that corresponding to the ($\alpha_{10}$, $\alpha_{50}$) quantile interval. The opposite occurs for low prices to a lesser extent. This asymmetry may be interpreted in terms of the uncertainty remaining for extreme prices. That is, when the prices are high, the model is not "confident" that they will be higher; and vice versa when they are low. Our interpretation is that the price formation, based on learned facts, makes the GBRT model emulate the behavior of actual agents, who submit bids with prices as low as possible. As a consequence, the market proper caps the maximum values of the price by rejecting the more expensive bids, what would suggest that the extreme prices would be rarer. This rationale would favor a negative skewness of the predictions of high prices, reflected in Fig. 5.

### 4.2. Electricity price analysis

As explained above, in addition to predicting the prices the more accurate 50-percentile GBRT model can also be used to investigate the sensitivity of the day-ahead price to changes in the explanatory variables. This can be done by analyzing the partial dependence—like in the well-known coefficient analysis of PCR models—and the variable importance.

We employed categorical variables related to month, week day, and hour of prediction to provide the GBRT with clues that served to single out hours for which the price prediction would be done (a total of 24 h on day $D + 1$). The results in Fig. 9 show that peak and off-peak hours and days are clearly identifiable. Importantly, they are valued. The first hours in the morning bring down the prices. It is at this time when in Spain the residual demand is sometimes negative, and wind power can be curtailed.

According to the GBRT model, it is the residual load forecast that has the largest impact on the price in Spain; provoking on average price deviations of around 5 €/MWh (Fig. 9a). In [16] the authors had presented a linear relationship between prices and load in Ireland. Differently, the 50-percentile GBRT model finds an S-shaped relationship, demonstrating a saturation of the price for extreme values of residual load forecast.

Again as expected, increasing wind power forecast drives down the price. However, it is noticeable from Fig. 9a that the price oscillations are the lowest for wind and intermediate for load forecast. The combination of both in the residual demand does not aggregate in a simple manner to yield the price sensitivity to the residual load, as shown in the panel.

In Fig. 9b, the non-linear effects are much more evident in the particular case of Morocco's link. So much so, that partial dependence analysis directly indicates that when the link is not used for energy export, the price is moved from the mean value by + 1 €/MWh. To some extent, it also occurs with imports from Morocco, revealing a sort of saturation effect that this type of analysis does not observe in the link

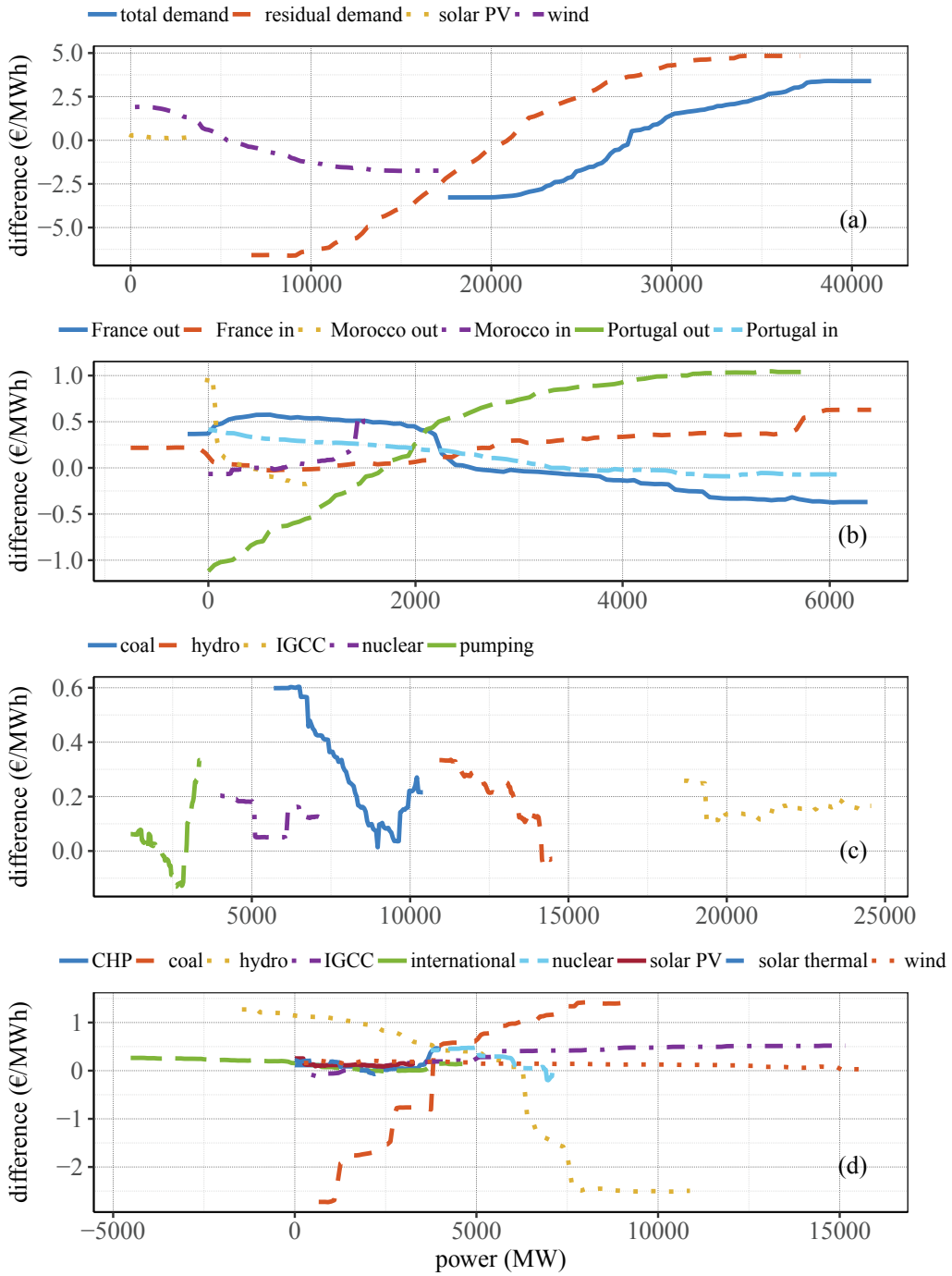



**Fig. 10.** Partial dependence (indeed the deviation, as in Fig. 9) between non-categorical predictors and the predicted day-ahead price. The four categories of predictors described in Section 2.2 are separately plotted. From top to bottom: forecasts, availability of international links, available dispatchable generation, and power generation at 11.00 a.m.

with Portugal. In the case of France, step-wise behaviors are observed at 0, 2 and 4.8 GWh.

Coal-based generators are undeniably the marginal units settling the price in Spain. This translates into the largest impact on price in their category. Interestingly, the price behavior is different depending on how these units are studied (compare Fig. 9c for availability and Fig. 9d for actual in-feed). The more units are available, the lower the price is expected to be (Fig. 9c) probably as a result of the perceived lower need for more expensive IGCC units in case of future low renewable in-feed. Regardless, the range of variation, 0.6 €/MWh, is relatively small. However, the impact derived from the observation of in-service power at 11.00 a.m. is opposite. The more units are in service at 11.00 a.m. the higher the price is expected to be; with a range of variation around 4.1 €/MWh.

Hydro power plants have the second largest impact on the price, according to the partial dependence analysis in panels c and d of Fig. 9. They account for 19.4% of the power generation in Spain, just behind IGCC (23.4%) and wind (21.9%) in 2017. (Coal-based power plants accounted for 9.1%.) Partial dependence analysis is useful to reveal that the number of in-service hydro power plants at market gate closure has far larger impact than IGCC and wind; which practically do not affect the price formation. Moreover, the non-linear characteristic of the price-hydro partial dependence serves to reveal that (i) the impact is larger when in-service plants add up between 6 and 8 GW, abruptly

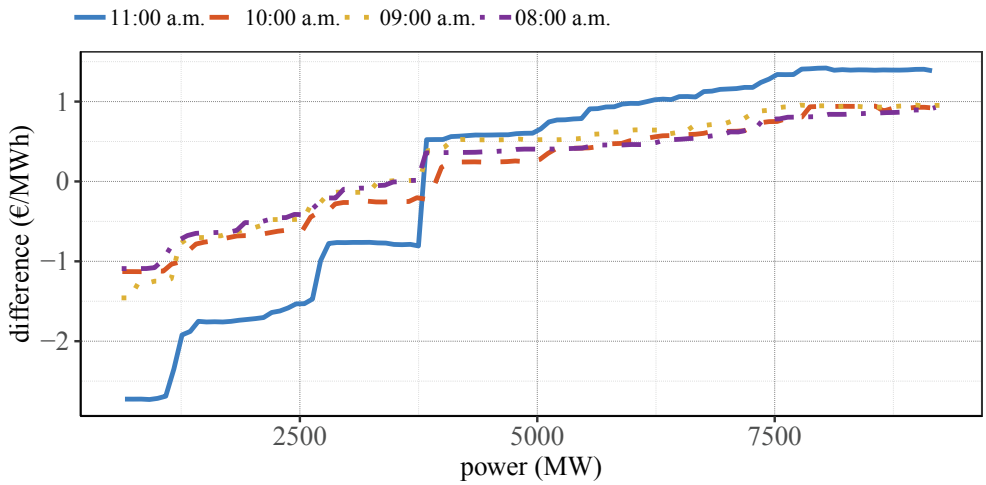



**Fig. 11.** Partial dependence between coal-based generation and the day-ahead price, depending on the values ahead of 11.00 a.m.

bringing down the price as the in-service power is larger; and (ii) that the effect of these plants on price above 8 GW is practically negligible.

Nuclear units are rarely marginal in Spain. Either through the analysis of unit availability (Fig. 9c) or real-time production (Fig. 9d), the fact that the capacity factor is close to 90% (in 2017 the installed power was 7.1 GW, with a production reaching 54.8 TWh) makes the effect on price variations to be remarkably small. It is noticeable, however, how the 50-percentile GBRT model captures the price variations due to changes in power. All nuclear units in Spain are of 1 GW. Most of the time, 4-GW nuclear power is in service, with the remaining two units sometimes being brought off-line. These rare events are captured by the model as shown in the third panel of Fig. 9 (which shows that the price is slightly reduced the more units are in service, with a noticeable drop in price when six units are in service) and in the fourth (which shows the step-wise evolution of the price in 1-GW intervals).

The step-wise relationship between power and price increments also occurs with coal units (see Fig. 9d). In Spain, coal-based generation accounts for 100 GW of installed power. Most units provide their biddings through a step-wise function of power blocks. An analysis of the size and price of these blocks using OMIE's datasets reveals that the bid pattern followed by each unit is quite cyclic and predictable, although large variations occur when some power is reserved to be sold in secondary markets (intraday adjustments or technical restrictions, mainly) in the expectation of more profitable returns. Being as they are marginal units, it seems reasonable to think that these events are best observed at the time of submitting the bids at 11.00 p.m. and projected into the future. That would add to explain the steep variations in the price up to 5 GW of in-service power at 11.00 a.m. Beyond that level, the price rises less steeply with the coal-based in-service power, but clearly it still reflects the addition of more expensive units.

We used the same type of analysis on coal-based generation to obtain Fig. 11, where it is shown the relation between price variation and in-service power as a function of the time ahead of the market gate closure. The partial dependence analysis reveals that the largest deviations are provoked by the value of in-service power at gate-closure, with the trajectory up to that point playing a secondary role, though also important. Interestingly, the effect of lags appears to repeat for high lags. The impact on price is almost the same after observing the in-service coal-based generation 10.00 and 8.00 a.m.

## 5. Conclusions

This paper has offered an overview about the usefulness of processing publicly accessible real-time data sources to make day-ahead predictions of the electricity price in Spain. The data source is a website where the Spanish Transmission System Operator (TSO) offers a graphical interface to a number of variables describing the power system state. Our approach has been to obtain selected variables from the website and define a prediction model for those specific explanatory variables. The outcome has been a simple quantile prediction model based on boosted gradient regression trees, which has shown good performance and has shown to provide insight into the price formation in the Spanish electricity market.

First, we have discussed the model performance. We have provided indicators about the accuracy and compared it with the results of a principal component regression model.

The calculated metrics show that our model produces remarkably low prediction errors when using the median as point prediction method (RMSE = 2.78 €/MWh, MAE = 1.94 €/MWh, and MAPE = 0.059).

Regarding the quantile regression, we have observed some interesting results. First, employing the percentile 50 as a point prediction of the price seems adequate, at least in the Spanish dataset analyzed. Additionally, our analysis has demonstrated a noticeable asymmetry in the prediction intervals. The distributional skewness has shown to be remarkable. Therefore, caution should be adopted when using symmetrical confidence bands around a central point prediction.

Interestingly, the quantile regression model also allows to inherently define prediction intervals, with a different interpretation of accuracy. Our results show that on average 90% of times the prediction error will not exceed 6.8 €/MWh.

Second, we have demonstrated that partial dependence analysis overcomes the black-box problem related to machine learning approaches, providing the basis for detailed analysis about the impact that individual predictors have on the price. Unlike the conventional linear trends obtained from the widely used multiple linear regression models, partial dependence has shown breaks or stepwise relationships, nonlinear trends, and extreme associations. Some of these have been easy to interpret, therefore could be included in this paper, and above all served as a confirmation of the usefulness of this type of analysis. Others cannot be so promptly explained, and as such they invite researchers to conduct more detailed analysis to unveil associations in the price formation.

## Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.apenergy.2019.01.213.

## References


[1] Zame KK, Brehm CA, Nitica AT, Richard CL, Schweitzer GD. Smart grid and energy storage: policy recommendations; 2018.
[2] Larsen EM, Pinson P, Leimgruber F, Judex F. Demand response evaluation and forecasting Methods and results from the EcoGrid EU experiment. Sust Energy Grids Networks 2017;10:75–83.
[3] McPherson M, Tahseen S. Deploying storage assets to facilitate variable renewable energy integration: the impacts of grid flexibility, renewable penetration, and market structure. Energy 2018;145:856–70.
[4] Chen Y, Trifkovic M. Optimal scheduling of a microgrid in a volatile electricity market environment: portfolio optimization approach. Appl Energy 2018;226:703–12.
[5] Liu Y, Yu N, Wang W, Guan X, Xu Z, Dong B, et al. Coordinating the operations of smart buildings in smart grids. Appl Energy 2018;228:2510–25.
[6] Prinsloo G, Dobson R, Mammoli A. Synthesis of an intelligent rural village microgrid control strategy based on smartgrid multi-agent modelling and transactive energy management principles. Energy 2018;147:263–78.
[7] e.sios Repository; 2018 [accessed: 2018-08-05]. < https://www.esios.ree.es/en/analysis/460 > .
[8] Elia; 2018 [accessed: 2018-12-08]. < http://www.elia.be/en/grid-data > .
[9] Fingrid; 2018 [accessed: 2018-12-08]. < https://www.fingrid.fi/en/electricity-market/ > .
[10] RTE; 2018 [accessed: 2018-12-08]. < http://clients.rte-france.com/lang/an/visiteurs/vie/courbes_methodologie.jsp > .
[11] EirGrid; 2018 [accessed: 2018-12-08]. < http://smartgriddashboard.eirgrid.com/ > .
[12] Svenska Kraftnat; 2018 [accessed: 2018-12-08]. < https://www.svk.se/en/national-grid/the-control-room/ > .
[13] Austrian Power Grid; 2018 [accessed: 2018-12-08]. < https://www.apg.at/en/markt/Markttransparenz > .
[14] Weron R. Electricity price forecasting: a review of the state-of-the-art with a look into the future. Int J Forecast 2014;30:1030–81.
[15] Singh N, Mohanty SR, Dev Shukla R. Short term electricity price forecast based on environmentally adapted generalized neuron. Energy 2017;125:127–39.
[16] Denny E, O'Mahoney A, Lannoye E. Modelling the impact of wind generation on electricity market prices in Ireland: an econometric versus unit commitment approach. Renew Energy 2017;104:109–19.
[17] Finn P, Fitzpatrick C, Connolly D, Leahy M, Relihan L. Facilitation of renewable electricity using price based appliance control in Ireland's electricity market. Energy 2011;36:2952–60.
[18] Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36(18?): 27–46.
[19] Peng L, Liu S, Liu R, Wang L. Effective long short-term memory with differential evolution algorithm for electricity price prediction. Energy 2018.
[20] Panapakidis IP, Dagoumas AS. Day-ahead electricity price forecasting via the application of artificial neural network based models. Appl Energy 2016;172:132–51.
[21] Lago J, De Ridder F, De Schutter B. Forecasting spot electricity prices: deep learning approaches and empirical comparison of traditional algorithms. Appl Energy 2018;221:386–405.
[22] Yang Z, Ce L, Lian L. Electricity price forecasting by a hybrid model, combining wavelet transform, ARMA and kernel-based extreme learning machine methods. Appl Energy 2017;190:291–305.
[23] Bento PM, Pombo JA, Calado MR, Mariano SJ. A bat optimized neural network and wavelet transform approach for short-term price forecasting. Appl Energy 2018;210:88–97.
[24] Ahmad MW, Mourshed M, Rezgui Y. Trees vs neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy Build 2017;147:77–89.
[25] Gaillard P, Goude Y, Nedellec R. Additive models and robust aggregation for GEFCom2014 probabilistic electric load and electricity price forecasting. Int J Forecast 2016;32:1038–50.
[26] Heredia FJ, Rider MJ, Corchero C. Optimal bidding strategies for thermal and generic programming units in the day-ahead electricity market. IEEE Trans Power Syst 2010;25:1504–18.
[27] Quan H, Srinivasan D, Khambadkone AM, Khosravi A. A computational framework for uncertainty integration in stochastic unit commitment with intermittent renewable energy sources. Appl Energy 2015;152:71–82.
[28] Nowotarski J, Weron R. Recent advances in electricity price forecasting: a review of probabilistic forecasting. Renew Sust Energy Rev 2018;81:1548–68.
[29] Nowotarski J, Weron R. Computing electricity spot price prediction intervals using quantile regression and forecast averaging. Comput Stat 2014;30:791–803.
[30] Bunn D, Andresen A, Chen D, Westgaard S. Analysis and forecasting of electricity price risks with quantile factor models. Energy J 2016;37:101–22.
[31] Mosquera-López S, Uribe JM, Manotas-Duque DF. Nonlinear empirical pricing in electricity markets using fundamental weather factors. Energy 2017;139:594–605.
[32] Hagfors LI, Bunn D, Kristoffersen E, Staver TT, Westgaard S. Modeling the UK electricity price distributions using quantile regression. Energy 2016;102:231–43.
[33] Hastie T, Friedman J, Tibshirani R. Boosting and additive trees. New York, NY: Springer; 2001.
[34] Anbazhagan S, Kumarappan N. Day-ahead deregulated electricity market price forecasting using neural network input featured by DCT. Energy Convers Manage 2014;78:711–9.
[35] Ziel F. Forecasting electricity spot prices using lasso: on capturing the autoregressive intraday structure. IEEE Trans Power Syst 2016;31:4977–87.
[36] Costello AB, Osborne JW. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Pract Assess Res Eval 2005;10:1–9.
[37] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7:21.
[38] REE's e.sios; 2018a [accessed: 2018-07-09]. < https://www.esios.ree.es/en > .
[39] e.sios Repository; 2018b [accessed: 2018-07-09]. < https://www.esios.ree.es/en/downloads > .
[40] e.sios Analysis; 2018c [accessed: 2018-07-09]. < https://www.esios.ree.es/en/analysis > .
[41] API e.sios Documentation. 2018d [accessed: 2018-07-09]. < https://api.esios.ree.es/ > .
[42] H2O.ai; 2018 [accessed: 2018-12-05]. < https://www.h2o.ai/products/h2o/ > .
[43] Østergård T, Jensen RL, Maagaard SE. A comparison of six metamodeling techniques applied to building performance simulations. Appl Energy 2018;211:89–103.
[44] Willmott CJ, Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res 2005;30:79–82.
[45] Friedman J. Greedy function appoximation: a gradient boosting machine. Source: Ann Stat Ann Stat 2001;53:1689–99.
[46] Tofallis C. A better measure of relative prediction accuracy for model selection and model estimation. J Oper Res Soc 2015;66:1352–62.
[47] Szkuta BR, Sanabria LA, Dillon TS. Electricity price short-term forecasting using artificial neural networks. Technical report 3; 1999.
[48] Keles D, Scelle J, Paraschiv F, Fichtner W. Extended forecast methods for day-ahead electricity spot prices applying artificial neural networks. Appl Energy 2016;162:218–30.
[49] Cuaresma JC, Hlouskova J, Kossmeier S, Obersteiner M. Forecasting electricity spot-prices using linear univariate time-series models. Appl Energy 2004;77:87–106.
[50] Ziel F, Steinert R, Husmann S. Efficient modeling and forecasting of electricity spot prices. Energy Econ 2015;47:98–111.
[51] Wang D, Luo H, Grunder O, Lin Y, Guo H. Multi-step ahead electricity price forecasting using a hybrid model based on two-layer decomposition technique and BP neural network optimized by firefly algorithm. Appl Energy 2017;190:390–407.
[52] Ghasemi A, Shayeghi H, Moradzadeh M, Nooshyar M. A novel hybrid algorithm for electricity price and load forecasting in smart grids with demand-side management. Appl Energy 2016;177:40–59.
[53] Raviv E, Bouwman KE, Van Dijk D. Forecasting day-ahead electricity prices: utilizing hourly prices. Energy Econ 2015;50:227–39.
[54] Uniejewski B, Weron R, Ziel F. Variance stabilizing transformations for electricity spot price forecasting. IEEE Trans Power Syst 2018;33:2219–29.
[55] Zhao JH, Dong ZY, Xu Z, Wong KP. A statistical approach for interval forecasting of the electricity price. IEEE Trans Power Syst 2008;23:267–76.
[56] Wan C, Xu Z, Wang Y, Dong ZY, Wong KP. A hybrid approach for probabilistic forecasting of electricity price. IEEE Trans Smart Grid 2014;5:463–70.