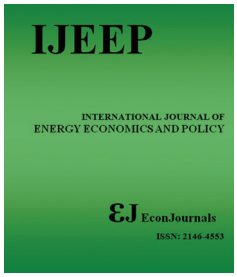


International Journal of Energy Economics and Policy

ISSN: 2146-4553

available at http: www.econjournals.com

International Journal of Energy Economics and Policy, 2021, 11(5), 66-77.

Electricity Price Fundamentals in Hydrothermal Power Generation Markets Using Machine Learning and Quantile Regression Analysis

Andrés Oviedo-Gómez[1]*, Sandra Milena Londoño-Hernández[1], Diego Fernando Manotas-Duque[2]

[1]School of Electrical and Electronic Engineering, Universidad del Valle, Cali, Colombia, [2]School of Industrial Engineering, Universidad del Valle, Cali, Colombia. *Email: oviedo.andres@correounivalle.edu.co

Received: 01 March 2021 **Accepted:** 10 June 2021 **DOI:** https://doi.org/10.32479/ijeep.11346

ABSTRACT

A hydrothermal power generation market is characterized by a strong dependence on water reservoir capacity and fossil fuel sources, which causes differences in generation marginal costs and high variability of the electricity spot price. Therefore, this study proposes an empirical approach to identify the price determinants and their effects on price dynamics. This paper presents two methodologies: a machine learning approach and a quantile regression analysis. The first method is used to validate the price determinants through a prediction process, and the second, the quantile regression, to identify the non-linear effects. The most important factors observed are total market demand, water reservoirs capacity for generation, and fossil fuel consumption. The results offer a new perspective about the market structure and spot price volatility.

Keywords: Electricity Prices, Hydrothermal Power Generation Markets, Machine Learning, Quantile Regression, Gaussian Process Regression
JEL Classifications: C22, Q41, Q43, Q47

I. INTRODUCTION

The different reforms in electricity markets defined electricity as a commodity, which can be sold, bought, and traded in a market (Berrie and Hoyle, 1985). However, its storage limitations make the market price shows characteristics such as seasonal patterns, high volatility, mean reversion, price spikes, and others (Girish and Vijayalakshmi, 2013; Huisman and Mahieu, 2003). Besides, modeling the price dynamic requires understanding its asymmetric distribution, high dispersion, and serial correlation (Ciarreta et al., 2011). Therefore, analyzing and predicting the spot prices is a challenge for academics and market agents.

On the other hand, the market structure and generation technologies are fundamentals factors in the price formation. Based on a particular case of a hydrothermal power generation market which presents: (i) significant differences in the marginal costs of the generation sector; (ii) a small renewable generation capacity; (iii) a strong dependence on exogenous variables as fossil fuel prices and climatology factors; and, where (iv) the risk and uncertainty are higher for market agents, it has been observed that these features cause further increased in price variability (Mosquera-López et al., 2017a; Fernández-Blanco et al., 2017; Cotia et al., 2019). Hence, it is relevant to recognize the determinants that explain the electricity price behavior in this market structure.

For this reason, the objective of this study was to identify the economic and technological fundamentals in the hydrothermal power generation market. Also, it was sought to evaluate fundamentals effects on spot price dynamic. For the empirical analysis, the Colombian electricity market was selected. Moreover, the methodology applied in this analysis was divided into two: a machine learning approach and a quantile regression analysis. First, a gaussian process regression (GPR) model was trained to

validate the determinants and compute the spot price prediction for the next 6 months of the dataset. This method identifies complex patterns in a large volume of data and reviews the data to predict future behavior (Castelli et al., 2020; Díaz et al., 2019; Gonzalez-Briones et al., 2019; Imani et al., 2020; Ribeiro et al., 2020). Second, a quantile regression model was fitted because it allows modeling electricity prices seasonality and quantifying the non-linear effects of determinants (Ma and Koenker, 2006; Maciejowska, 2020; Mosquera-López et al., 2017b; Uribe and Guillen, 2020).

According to Aggarwal et al. (2009) and Girish and Vijayalakshmi (2013), the spot price determinants were grouped into five categories: (i) market characteristics, (ii) fundamental factors, (iii) operation factor, (iv) strategic factors, and (v) historical factors. In the first group, it was identified variables such as energy supply and demand, electricity exports/imports, market-clearing quantity, and energy policy (Deng and Oren, 2006; Mandal et al., 2007; Mosquera-López and Nursimulu, 2019; Zhang et al., 2008). In the second group, the fundamental factors considered were price volatility, fuel price, weather factors, and hydrological conditions. By contrast, operational factors describe fundamentals as a system load rate, electricity production (deficit/surplus), energy sources (nuclear, hydric, or thermal), line status and limits, and power transmission costs (He et al., 2010; Rodriguez and Anders, 2004; Zhang et al., 2008). Meanwhile, strategic factors correspond to energy purchasing agreements, bilateral contracts, bidding strategy, and market design (Crespo-Cuaresma et al., 2004; Kian and Keyhani, 2001; Rodriguez and Anders, 2004). Finally, in the fifth group, it has been identified that past observations of variables as demand and supply, hydric reserves, and electricity price affect the present spot price dynamic (Ciarreta et al., 2011; Mandal et al., 2007).

However, and based on the power generation structure selected, the results of the empirical application described that total market demand, water reservoirs capacity for generation, and fossil fuel consumption, are the most relevant determinants of the spot price. Also, this paper provides a new contribution in terms of market structure analysis and a new perspective of the spot price distribution.

The paper is structured, after section 1, as follows: section 2, it is described the structure of the Colombia electricity market. Section 3 presents the empirical methodologies, and, in section 4, the dataset is described. In section 5, the results are reported, and section 6 presents the conclusions.

2. COLOMBIAN ELECTRICITY MARKET

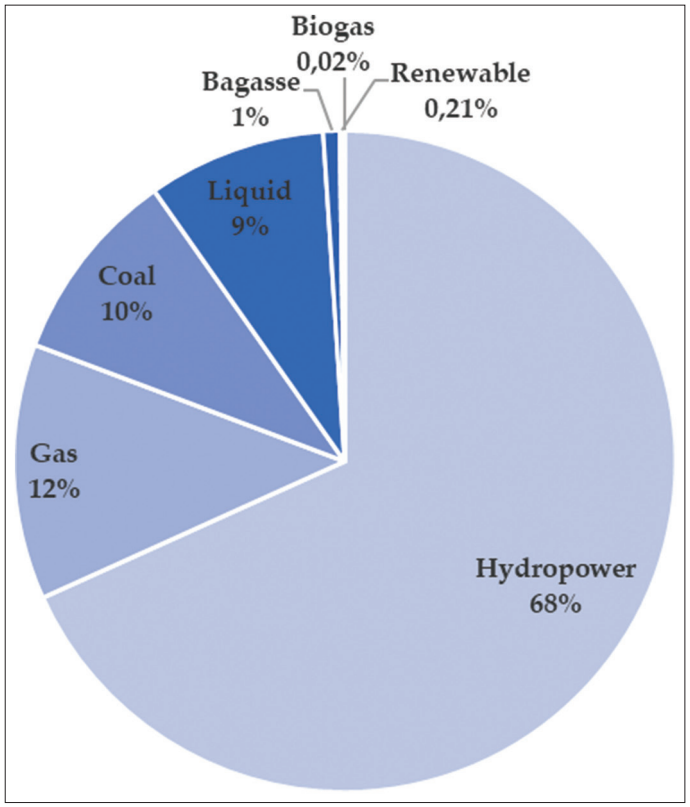
Since 1990, the Colombian energy sector has presented relevant reforms. García et al. (2011) described that the liberalization process allowed an improvement in the electricity market by introducing competition in different sectors, and hence, abolish the limitations of the vertical structure. Besides, the wholesale energy market (WEM) was created under a regulatory framework, and its operation is through a trade spot structure. However, the electricity sector presents limitations such as a low generation capacity and high demand, which do not allow structuring a competitive market, and electricity prices cannot capture the relationships between the supply and demand (Barrientos et al., 2012).

On the other hand, Colombia is part of a region with a lot of hydric sources. According to International Energy Agency (IEA) statistics, in 2018, approximately 86% of power generation in Central and South America was through hydric and thermal generation. Therefore, Colombia is part of these hydrothermal generation systems, where hydroelectric power generation represents 68% and thermal power generation (gas, coal, and liquid) 31% (Figure 1). While, renewable sources do not have a representative value in the power generation matrix (0.21%).

Due to hydrothermal power generation dependence, the Colombian electricity sector presents a high vulnerability by two exogenous factors: El Niño–Southern Oscillation (ENSO) and energy fossil price fluctuations. Figure 2 shows the daily spot price dynamic for the period 2000-2019, and significant effects of ENSO were observed in four periods during 2003 and 2014; however, the most important shock was observed between 2015 and 2016, where the price reached a maximum peak, and the gas prices increased considerably. Besides, the thermal generation sector did not have an economic guaranty to cover the demand[1]; hence, the state intervened in the market to avoid rationing (Botero-Duque et al., 2016; Montes, 2018).

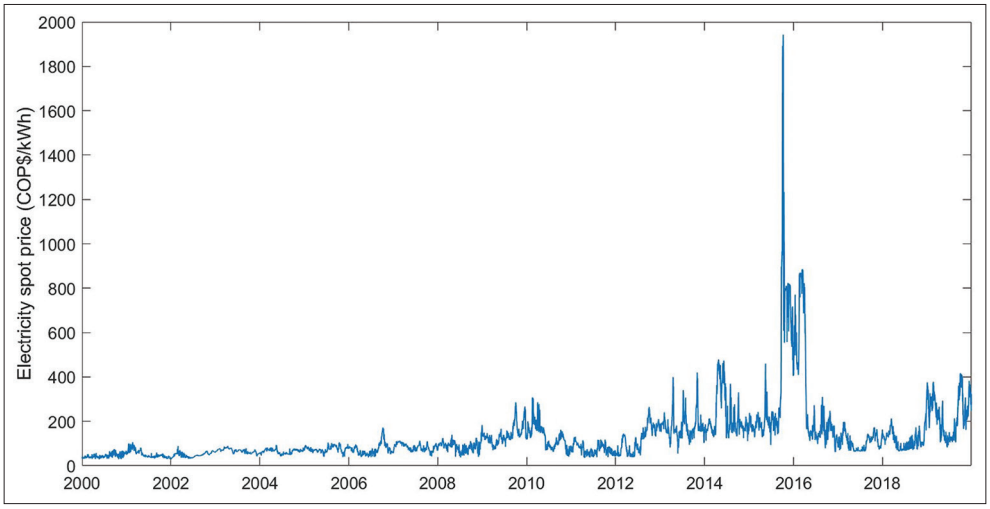
1 Thermal generation is a backup source for hydropower generation in two specific moments: high demand or low water reservoir levels.

Figure 1: Power generation net capacity by technology for January 2020



Source: XM information system.

Figure 2: Electricity spot price dynamic for the period 2000-2019



Source: XM information system.

According to Castaño and Sierra (2012), Díaz-Contreras et al. (2014), Lira et al. (2009), and Quintero-Quintero and Isaza-Cuervo (2013), the spot price is related to weather changes, fossil fuels to the thermal power generation, and electricity demand and supply. Likewise, power transmission failures, energy policy, or agent strategies are significant. Finally, Mosquera-López et al. (2017b) described that differences in the marginal costs are forwarded into the spot price dynamics and, consequently, increases the risk to agent decision making.

3. METHODOLOGY

Two approaches were considered to analyze the fundamentals of the electricity spot price in a hydrothermal power generation market. First, a machine learning approach was used, through a GPR, to fit a multivariable model to predict daily electricity price and validate the importance of variables considered; second, a quantile regression model was fitted to evaluate the effects of these predictors on the electricity price dynamic.

3.1. Gaussian Process Regression Models

According to Rasmussen and Williams (2006), and The Mathworks (2020), the GPR models are nonparametric kernel-based models of supervised learning, used for regression analysis and probabilistic classification. These models capture uncertainty and allow predictions where the data have unknown distributions. Besides, the GPR is a powerful method to perform Bayesian inference, and it is much better when the availability of the data is a problem (Aye and Heyns, 2017; Gonzalez-Briones et al., 2019).

A training set is defined as $\{(x_i,y_i);i=1,2,\ldots,n\}$, where $x_i \in R^d$ and $y_i \in R$, and have an unknown distribution. Based on a linear regression model, a GPR model predicts the response variable by introducing latent variables, $f(x_i)$,i=1,2,…,n, from a gaussian process (GP), and explicit basis function, h.

A GP is defined by its mean function, $m(x)$, and covariance function, $k(x,x')$. If $\{f(x),x \in R^d\}$ is a GP, then $E(f(x))=m(x)$ and $cov[f(x),f(x')\}=E[\{f(x)-m(x)\}\{f(x')-m(x')\}]=k(x,x')$. Therefore, it considers the following model:

$$h(x)^T \beta + f(x), \tag{1}$$

Where $f(x) \sim GP(0,k(x,x'))$, i.e. $f(x)$ is zero mean GP with covariance function $k(x,x')$. Besides, $h(x)$ is a set of basis functions that project the input x into a new p-dimensional feature space vector (R^p) and β is a $px1$ dimension vector of basis function coefficients. This is a representation of GPR model and the response variable can be described as:

$$P(y_i|f(x_i),x_i) \sim N(y_i|h(x_i)^T \beta + f(x_i),\sigma^2). \tag{2}$$

Therefore, a GPR model is a probabilistic model. Furthermore, the GPR model is nonparametric model because of the observation x_i has a latent variable $f(x_i)$.

The joint distribution of latent variable $f(x_1),f(x_2),f(x_3),\ldots,f(x_n)$ in the GPR model is $P(f|X) \sim N(f|0,K(X,X))$, close to a linear regression model, where $K(X,X)$ is the covariance function and can be parametrized by a set of kernel parameters, θ. Hence, $k(x,x')$ is written as $k(x,x'|\theta)$ to explicitly indicate the dependence on kernel parameters.

3.1.1. Kernel function options

The kernel parameters are based on the signal standard deviation σ_f and the characteristic length scale σ_l. The characteristic length scales define the distance between the input values x_i and response values to become uncorrelated. The standard deviation and the characteristic length scale must be greater than zero, given $\theta_1=log\sigma_l$ and $\theta_2=log\sigma_f$.

The following four built-in kernel function with the same length scale were considered:

- Rational quadratic Kernel

$$k(x_i,x_j|\theta) = \sigma_f^2 \left(1+\frac{r^2}{2\alpha\sigma_l^2}\right)^{-\alpha}, \tag{3}$$

where σ_l is the characteristic length scale, α is the positive-valued scale-mixture parameter, and $r=\sqrt{\left((x_i-x_j)^T(x_i-x_j)\right)}$ is the Euclidean distance between x_i and x_j.

- Squared exponential kernel

$$k\left(x_i,x_j|\theta\right)=\sigma_f^2 e^{\left[-\frac{1}{2}\frac{\left(x_i-x_j\right)^T\left(x_i-x_j\right)}{\sigma_l^2}\right]}, \quad (4)$$

where σ_l is the characteristic length scale and σ_f is the signal standard deviation.

- Matern 5/2

$$k\left(x_i,x_j|\theta\right)=\sigma_f^2\left(1+\frac{\sqrt{5}r}{\sigma_l}+\frac{5r^2}{3\sigma_l^2}\right)e^{\left[-\frac{\sqrt{5}r}{\sigma_l}\right]}, \quad (5)$$

- Exponential

$$k\left(x_i,x_j|\theta\right)=\sigma_f^2 e^{\left[-\frac{r}{\sigma_l}\right]}, \quad (6)$$

where σ_l is the characteristic length scale and r is the Euclidean distance between x_i and x_j.

3.1.2. Parameter estimation

To estimate the parameters β, θ, and σ^2 of a GPR model, the likelihood $P(y|X)$ must be maximized as a function of parameters:

$$\hat{\beta},\hat{\theta},\hat{\sigma}^2=\underset{\beta,\,\theta,\,\sigma^2}{\operatorname{argmax}}\, logP(y\mid X,\beta,\theta,\sigma^2). \quad (7)$$

Because, $P(y|X,\beta,\ \theta,\ \sigma^2)=N(y|H\beta,K(X,X|\theta)+\sigma^2\ I_n)$, the marginal log-likelihood function is as follows:

$$\begin{aligned} logP(y\mid X,\beta,\theta,\sigma^2)&=-\frac{1}{2}\left(y-H\beta\right)^T\\ &\left[K\left(X,X|\theta\right)++\sigma^2 I_n\right]^{-1}\left(y-H\beta\right)-\frac{n}{2}log2\pi\\ &-\frac{1}{2}\log\left|K\left(X,X|\theta\right)+\sigma^2 I_n\right|, \end{aligned} \quad (8)$$

where, H is the vector of explicit basis functions, and $K(X,X\,|\,\theta)$ is the covariance function. To estimate the parameters, first, $\hat{\beta}(\theta,\sigma^2)$ is determined and its estimation is used to compute the β-profiled likelihood. Second, the β-profiled log-likelihood is given by $logP(y\mid X,\hat{\beta}(\theta,\sigma^2),\theta,\sigma^2)$, where it maximizes the β-profiled log-likelihood over θ, σ^2 to find their estimates.

Finally, during the estimation process, principal component analysis (PCA) was applied to avoid multicollinearity and dimensionality problems.

3.1.3. Response variable forecast

To predict a value of a response variable y_{new}, given a new input vector x_{new}, and the training data, it is defined the density $P(y_{new}|y,X,x_{\text{new}})$ by conditional probabilities:

$$P\left(y_{new}|y,X,x_{new}\right)=\frac{P\left(y_{new,}\,y|X,x_{new}\right)}{P\left(y|X,x_{new}\right)}. \quad (9)$$

To find the joint density in the numerator, it is necessary to introduce the latent variables f_{new} and f corresponding to y_{new}, and y, respectively. Thus, it is possible to use the joint distribution for y_{new}, y, f_{new}, and f to compute (9). The GP models assume that each response only depends on the corresponding latent variable f_i and the feature vector x_i.

After we found the density $P(y_{new}|y,X,x_{new})$, the expected value of prediction y_{new} at a new point x_{new}, given y, X, and parameters β, θ, σ^2 is:

$$E\left(y_{new}|y,X,x_{new},\beta,\ \theta,\ \sigma^2\right)=h(x_{new})^T\beta+K\left(x_{new}^T,X|\theta\right)\alpha, \quad (10)$$

where, $\alpha=\left(K\left(X,X|\theta\right)+\sigma^2 I_n\right)^{-1}(y-H\beta)$.

3.1.4. Performance indicators

To check the GPR model performance, different calibration metrics were used such as root mean square error (RMSE), R-squared (R^2), and mean absolute error (MAE). These metrics are described in the following:

- RMSE

$$RMSE=\sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2\right)}, \quad (11)$$

- R2

$$R^2=1-\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\bar{y}_i\right)^2}, \quad (12)$$

- MAE

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_i\right|, \quad (13)$$

where n is the number of observations, y_i is the i-th observed value, and $\hat{y}_i$ is the i-th predicted value. For RMSE and MAE lower values are desired, and for R^2, a closest value to one shows a better performance. Besides, the performance metrics of the estimated GPR model were compared with two supervised learning models: Support vector machines (SVM) and Tree-based methods. The performance metrics are described in the result and discussion section.

3.2. Quantile Regression Approach

The quantile regression is a semi-parametric approach, with high flexibility that captures the stochastic relationship between variables, allows consistent estimation in non-Gaussian environmental, and requires a minimal distributional assumption on the data generating process (Koenker, 2004; Ma and Koenker, 2006; Uribe and Guillen, 2020). To describe the quantile regression model, a linear regression model was assumed, where the response variable $Y_{i,t}$ represents the electricity spot prices and is related to a set of explanatory variables or fundamentals in a matrix $X_{i,t}$. Following Koenker and Bassett (1978), the quantile regression model can be written as:

$$Q_q\left(Y_{i,t}|X_{i,t}\right)=X_{i,t}^{'}\beta_i^q, \quad (17)$$

where, $Y_{i,t}$ is a (T*x*1) vector, with T denoting the number of observations (t=1,2,3,…T). Besides, the matrix $X_{i,t}^{'}$ of dimensions (T*x*d), has (d-1) predictors that also includes a constant, and β^q is a (d*x*1) vector of unknown parameters for each quantile q, $q\in(0,1)$. The regression coefficients $\hat{\beta}^q$ of the quantile q were estimated as a solution to the following minimization problem:

$$\min_{\beta_i^q}\frac{1}{T}\sum_{t=1}^{T}\left[q-I\left(Y_{i,t}<X_{i,t}^{'}\beta_i^q\right)\right]\left[Y_{i,t}-X_{i,t}^{'}\beta_i^q\right], \quad (18)$$

where,

$$I\left(Y_{i,t}<X_{i,t}^{'}\beta_i^q\right)=\begin{cases}1, Y_{i,t}<X_{i,t}^{'}\beta_i^q\\ 0, otherwise\end{cases}, \quad (19)$$

$Y_{i,t}$ is defined as in equation (17) and must be computed in separate regressions for each i, i=1,…N. According to Mosquera-López et al. (2017b), and Uribe and Guillen (2020), the quantile regression is a special case of the least absolute deviation estimator (LAD), that allows robust estimations when the data present heavy tails as for electricity spot prices.

4. DATA

The fundamentals of spot price are determined by the generation technologies. For example, in Central and South America, the generation is based, principally, on hydroelectric and thermal power sources. In this cases, different studies have described the following determinants: demand, hydrology changes, fossil fuel price variation, investment decisions making, the structure of the transmission system, and agent strategies (Barria and Rudnick, 2011; Barrientos-Marín and Toro-Martínez, 2017; Blazsek and Hernández, 2018; Samudio-Carter et al., 2019; Vaca et al., 2019; Xavier et al., 2016). Therefore, the first database contained variables such as (i) total demand: real, commercial, and National Interconnected System (NIS); (ii) reservoir levels: daily volume in percentage and generation capacity; (iii) climatology factors as quantity of water that fuel reservoirs; and (iv) fuel fossil consumption: gas, coal, fuel oil, and kerosene. On the other hand, variables as the bilateral bidding price, electricity imports/exports, or the price regulatory policies were not selected due to the spot price is contained in their structures or missing observations were identified.

According to the variables described, finally, the correlation analysis was used to select the spot price determinants. Besides, considering the capacity of generation (Figure 1), the volume of water available in the reservoirs and the consumption of fossil fuels from two of the most important sources, gas and coal, were selected. Also, NIS demand was chosen because this variable is calculated based on the net generation of the plants. These variables were chosen due to they allow the structure of a parsimonious model characterized by describing a classic supply and demand model. The dataset applied in this research represents the market structure and seeks to explain the spot price dynamic. Table 1 shows the variables, specifying data source and units.

In summary, the database is a balanced panel composed of daily data that starts in August 2009 and ends in December 2019. The period was determined because of the availability of data with no methodological changes, and the current supply scheme for the generation sector is included (Creg 051 of 2009, article 10). Likewise, 2020 data were not selected because regulated and non-regulated demand decreased by 4.2% and 12.9%, respectively, during the first quarterly by the SARS-CoV-2 (COVID-19) pandemic (Vidal et al., 2020).

Table 2 reports summary statistics and unit root test (augmented Dickey-Fuller - ADF) of the variables and Figure 3 describes their dynamics during the sample period. The spot price presents a high variability and dispersion, especially in the last quartile due to ENSO effects during 2015 and 2016, where the price increased to 1943 COP$/kWh. Then again, the demand has a dynamic growth and shows a correlation of 0.26 with the price, which is positive and weak, despite the demand is a significant price determinant. Regarding water volume, it was observed a high variability by seasonal patterns and a negative correlation with price. Likewise, gas and coal are sources used to supply the demand when the hydropower system presents any limitation. Hence, these variables have a high dispersion in the last quartiles

2. COP is the representative sign of the Colombian peso.

Table 1: Data description

Variable	Description	Units	Source
Spot price	Daily electricity spot price	COP[2]$ / kWh	XM information system
Demand	Total demand with energy losses	MW/h	XM information system
Water volume	Reservoirs capacity for hydropower generation	Percentage or GW/h	XM information system
Gas	Gas quantity consumption	MBTU	XM information system
Coal	Coal quantity consumption	MBTU	XM information system

Source: au Thor's construction

Table 2: Summary and ADF test for selected variables

Statistical parameters	Spot price	Demand	Water volume (GW/h)	Gas	Coal
Mean	184.47	173571	10527	231712	121932
Std. Dev.	166.96	18334.57	2122.958	94968.62	71163.86
Minimum	35.36	115438	5777	63336	0
25th percentile	97.93	160645	9022	155753	61913
50th percentile	146.88	173729	10712	211287	117501
75th percentile	194.68	188247	12303	289651	174210
Maximum	1942.69	217021	14502	543258	356137
Spot price correlation	-	0.23	−0.26	0.45	0.60
t-ADF	−4.63***	−8.53***	-3.70**	−4.10***	−6.06***

** and *** indicates that null hypothesis of a unit root is rejected at 5% and 1% level, respectively.

Source: Authors' analysis

and a significant and positive correlation with the price. Finally, ADF test was computed, and the result shows evidence against the presence of unit root in the variables for a 1% and 5% level of confidence. Therefore, the variables do not require stationary transformation before the estimation.

4.1. Determining, Training and Testing Set for Machine Learning Approach

Figure 4 summarizes the machine learning methodology through the variable set described. First, the dataset was imported from XM Information System, explored, and processed to find their descriptive statistics and identify their characteristics. In general, the variables did not transform, except for the spot prices due to outliers observed during the 2015-2016 period. Spot price outliers were filled through the Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) to avoid their effects in the prediction process and a possible overfitting.

Second, a training set is used to train the model, while a validation set is used to evaluate the model performs with the dataset by the performance indicators, and a final test is used to confirm the model specification and identify overfitting. Therefore, hold-out method was used to divide the dataset into three parts: train (65%), validation (15%), and test (20%) sets[3]. In this process stage, the response of the variables and their predictors were defined. According to the GPR model described in equation (2), we can write it in vector form:

$$P(y|f,X) \sim N\left(y|H\beta+f,\sigma^2 I\right), \tag{20}$$

where, the response variable y is the spot prices and the vector X has the fundamentals: demand, water volume, and gas and coal consumption.

Third, the best models were identified through performance indicators and the prediction of the daily spot price for the period August 2019-December 2019 was implemented.

4.2. Determining Quantile Regression Model

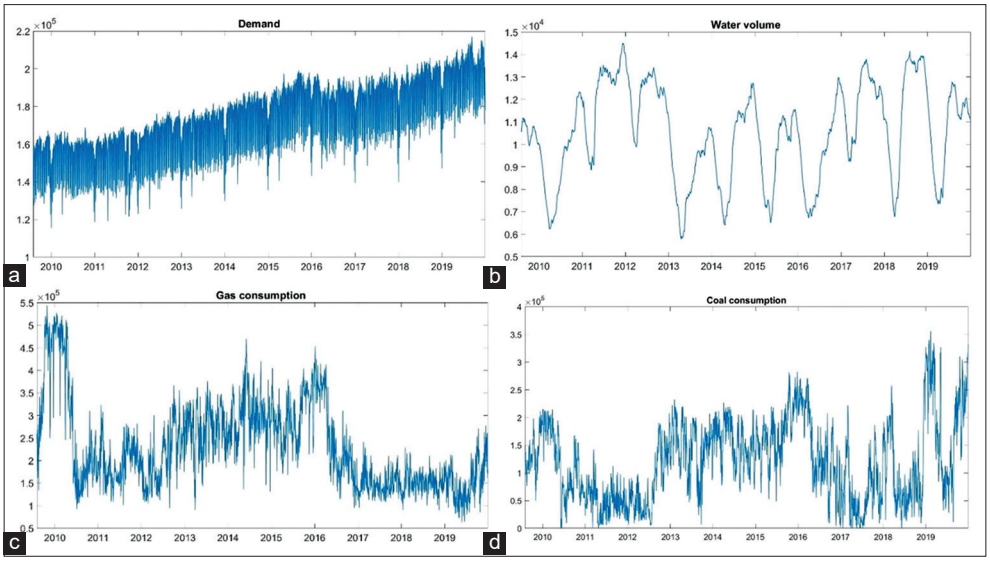
Based on equation (17), the linear quantile regression model can be written as a function of the response variable and their predictors:

$$Q_q\left(P_{i,t}\right) = \beta_{i,1}^q + \beta_{i,2}^q D_t + \beta_{i,3}^q W_t + \beta_{i,4}^q C_t, \tag{21}$$

where, P_t is the response variable, spot price, while D_t is the demand, W_t is water volume, and C_t is the total gas and coal consumption. For estimating the quantile regression model, the period August

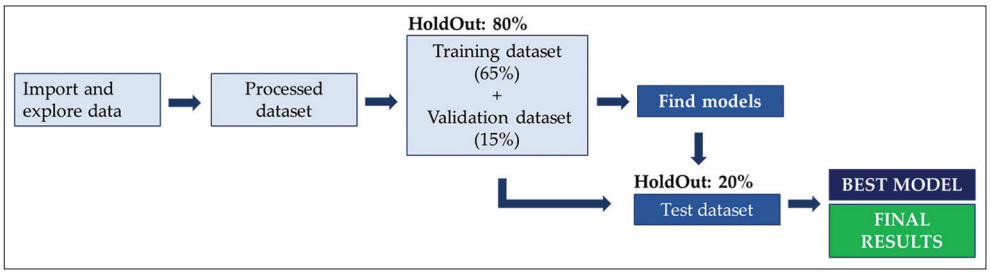
3 For train, validation, and test sets, the period August 2009-July 2019 was used.

Figure 3: Evolution of fundamental variables for August 2009-December 2019. (a) The National Interconnected System (NIS) demand in MW/h; (b) water volume or reservoir capacity in GW/h; (c) Gas consumption for generating in MBTU; (d) Coal consumption for generating in MBTU.



Source: Author's construction.

Figure 4: Summary for machine learning methodology



Source: Authors' analysis.

2009-December 2019 was used and the natural logarithms were computed to interpret the coefficients as elasticities.

5. EMPIRICAL RESULTS AND DISCUSSION

The main findings are presented below for the variables and timespan selected. First, the machine learning training results and performance metrics are described. Then, the daily spot price prediction is shown. Second, in this section, the results of the quantile regression analysis are described to identify the effects of the main determinants on the spot price.

5.1. Machine Learning Results

The performance of the GPR model fitting is assessed using RMSE, R^2, and MAE metrics[4]. Besides, the GPR model was compared with the support vector machine, which is categorized as a supervised learning method for the application of regression and classification. This method is based on determining hyperplanes that maximize the margin between classes (Gao et al., 2008). The following SVM kernels were used:

- Quadratic
- Cubic
- Gaussian: fine, medium, and coarse.

On the other hand, tree-based methods were considered due to their fast for fitting and prediction, low memory usage, and ease of interpretation. Therefore, the models used were: fine, medium, and coarse, for tree regression.

Besides, the training process is computed through PCA. Therefore, it was observed that models were estimated through the first two principal components due to these factors explained 98% of the variation of the selected determinants.

Tables 3-5 describe the metric performance for different fitting models and the kernels selected. Based on all performance metrics, the results show that the GPR Exponential performs better. In general, good performance was observed for the GPR models because the metrics for the three sets used were similar, in contrast to the SVM models that present a significant difference in the RMSE between the train and the other two datasets. Therefore, this leads us to conclude the possibility of overfitting in the SVM models. However, the SVM models presented a similar performance in validation and test sets in the MAE metric, this could suggest that the models still have a good predictive process. Then again, some differences were observed in tree regresion metrics; but the Medium and Coarse models presented a similar RMSE and MAE during the train, validation, and test sets. Finally, the R^2 shows the percentage of the dependent variable variation that explain by the model, but some of these models are not linear, so the use of this indicator may be subject to criticism (Díaz et al., 2019).

According to Barrientos-Marín and Toro-Martínez (2017), another performance indicator is the mean absolute percentage error (MAPE). This metric describes the relative absolute deviation in per unit value. For each of the GPR models, the MAPE is 21%, for SVM models, the lowest MAPE is 21% for fine and medium Gaussian kernels through the test dataset. Likewise, the Coarse for tree regression has a MAPE equal to 22%.

Table 3: Metrics performance for GPR models

Model's stages	RMSE	R^2	MAE
Kernel: Rational quadratic			
Train	47.57	0.78	32.21
Validation	43.42	0.81	30.11
Test	43.40	0.79	29.48
Kernel: Squared Exponential			
Train	48.20	0.77	32.21
Validation	43.52	0.81	30.16
Test	43.51	0.79	29.55
Kernel: Matern 5/2			
Train	47.82	0.78	32.40
Validation	43.43	0.81	30.10
Test	43.22	0.79	29.39
Kernel: Exponential			
Train	44.45	0.81	30.16
Validation	43.17	0.82	29.79
Test	42.55	0.80	28.94

Source: Authors' analysis

Table 4: Metrics performance for SVM models

Model's stages	RMSE	R^2	MAE
Kernel: Quadratic			
Train	58.19	0.67	39.55
Validation	54.66	0.71	37.13
Test	54.26	0.67	36.54
Kernel: Cubic			
Train	53.72	0.72	36.33
Validation	48.82	0.76	33.50
Test	50.32	0.72	34.29
Kernel: Gaussian fine			
Train	49.12	0.76	30.38
Validation	45.07	0.80	29.76
Test	44.94	0.78	28.82
Kernel: Gaussian medium			
Train	51.09	0.74	33.67
Validation	45.87	0.79	31.14
Test	46.71	0.76	30.95
Kernel: Gaussian Coarse			
Train	61.33	0.63	39.78
Validation	59.12	0.65	38.26
Test	56.96	0.64	36.39

Source: Authors' analysis

Table 5: Metrics performance for tree regressions

Model's stages	RMSE	R^2	MAE
Fine model			
Train	36.32	0.87	22.96
Validation	51.97	0.73	34.76
Test	50.37	0.72	33.40
Medium model			
Train	43.24	0.82	28.66
Validation	43.69	0.79	31.17
Test	44.76	0.78	30.82
Coarse model			
Train	46.62	0.79	31.19
Validation	43.69	0.81	29.65
Test	45.99	0.77	30.76

Source: Authors' analysis

4 RMSE and MAE metrics value in COP$/MWh.

In summary, it was observed better performance metrics for the GPR models, especially the GPR exponential model. These models provide predictions for a given spectrum and a predictive distribution that allows computing the first and second moments: the mean and the standard deviation. Likewise, the kernels that provide rankings of the input variables or variance estimation of the data noise. Hence, the GPR models offer an alternative for analyzing a variable that presents mean reversion, spikes, and seasonal patterns.

Therefore, the daily prediction was computed through this model for the period from August 2019 to December 2019 (Figure 5). The dynamics of the spot price generated by the selected predictors were observed and it was conclude that the model allows a good approximation for lower prices, i.e., under 250 COP$/kWh. However, the prediction has not reached the true value for high prices, especially during the period from September to November. Barrientos-Marín and Toro-Martínez (2017) described for the Spanish market, an asymmetry response between the high and low prices. When the price is high, the model does not believe that prices will be higher. Likewise, when the price is low, the model is not confident that prices will be lower. Therefore, the authors explained that their model could capture the agent behavior, who submit bids with low prices to compete. Nevertheless, Weron (2014) and Ziel (2016) described there is not a standard structure for the electricity markets. Hence, it is not possible to make a comparison between markets and performance metrics for machine learning approach.

By contrast, the average spot price during July 2019 was 123.57 COP$/kWh, and during October 2019, the price reached an average of 390.4 COP$/kWh. A reduction in hydric sources during August and September could explain the high price increase; however, the water reservoirs had a percentage of 74% and 67% in August and September, respectively. Besides, the water reservoir percentage in October and November was approximately 69%. Therefore, the generation concentration index or an oligopolistic indicator must be considered because the hydropower generation tries to make speculations when the water sources decrease and, thus, increase the price in the following months (Aggarwal et al., 2009; Zhang and Luh, 2005).

5.2. Quantile Regression Results

For estimating the quantile regression model, the complete sample was used: August 2009-December 2019. Likewise, the response variable was not transformed by outliers due to quantile regression models are robust to these data and according to Uribe and Guillen (2020) the financial time series presents crises and booms with high or low observations.

Figure 6 describes the spot prices' quantile against the corresponding fraction of data. A low spot price was observed for the lower quantiles, approximately equal to 35 COP$/kWh, and around 147 COP$/kWh for the median price. From the lower to higher quantiles, a smooth increase was identified; however, after 85% quantile, the price presents a sharp peak related to exogenous effects during 2015 and 2016.

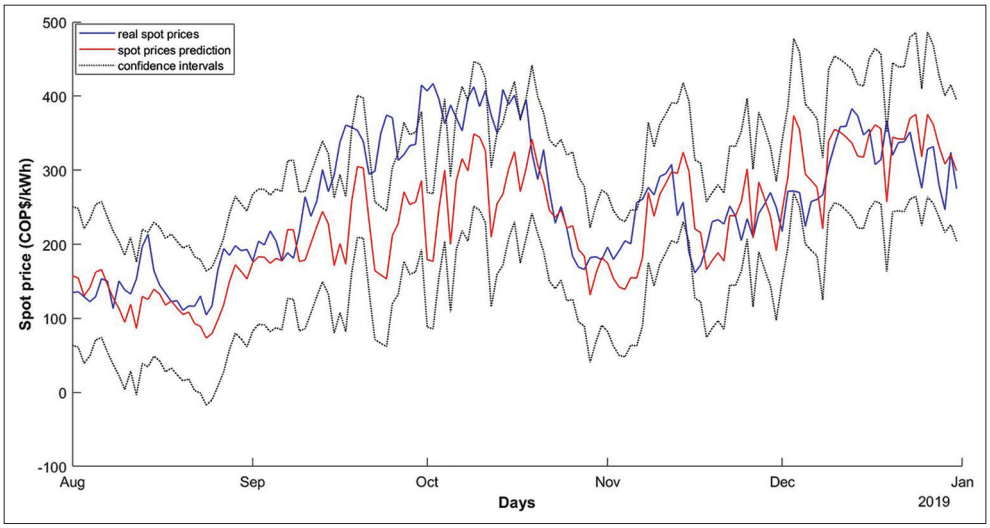
The linear model described in (21) was estimated for different percentiles of the distribution of electricity prices, i.e. from the 10th to the 90th percentile. Furthermore, the gas and coal consumption were added to analyze the proportion of fossil fuel consumption due to these two variables are around 22% of total generation capacity. The main results are summarized in Figure 7 and the quantile regression coefficients are presented in the Appendix A.

5.2.1. Effects of the determinants variables of the electricity spot price for different percentiles

5.2.1.1. Demand effects

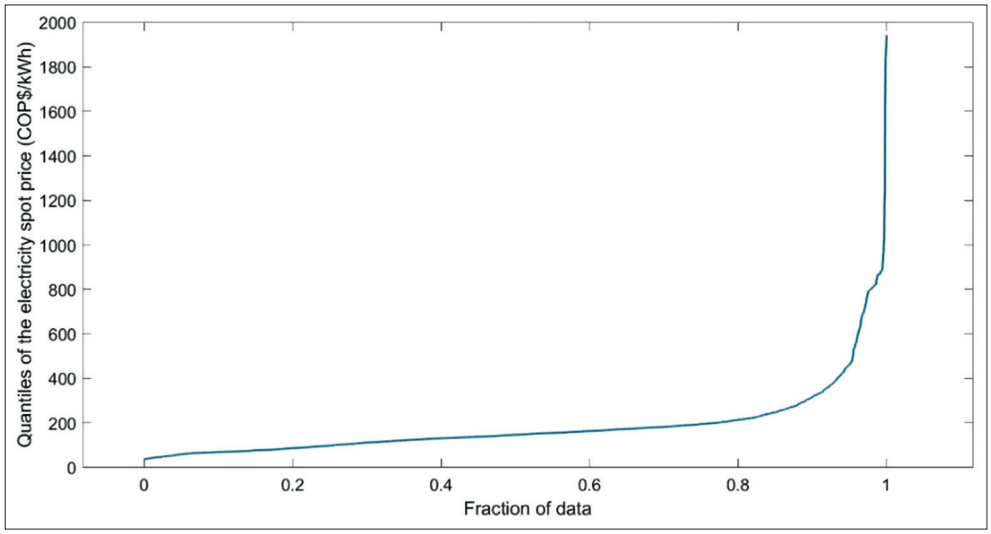
The sensitivity to changes in demand is positive and significant statistically, but its effects vary over the different spot price quantiles. In the 10th percentile, where the price is low, the demand presents a high impact, e.g. for a demand variation of 1%, the price variation is approximately 2%. However, around the 20th to the 50th percentile, the demand impact decreases significantly. For

Figure 5: Electricity spot price daily prediction for August 2019-December 2019. The continues blue line is the real spot price for the sample. The dotted red line is the spot price prediction, and the dotted black lines are the prediction intervals



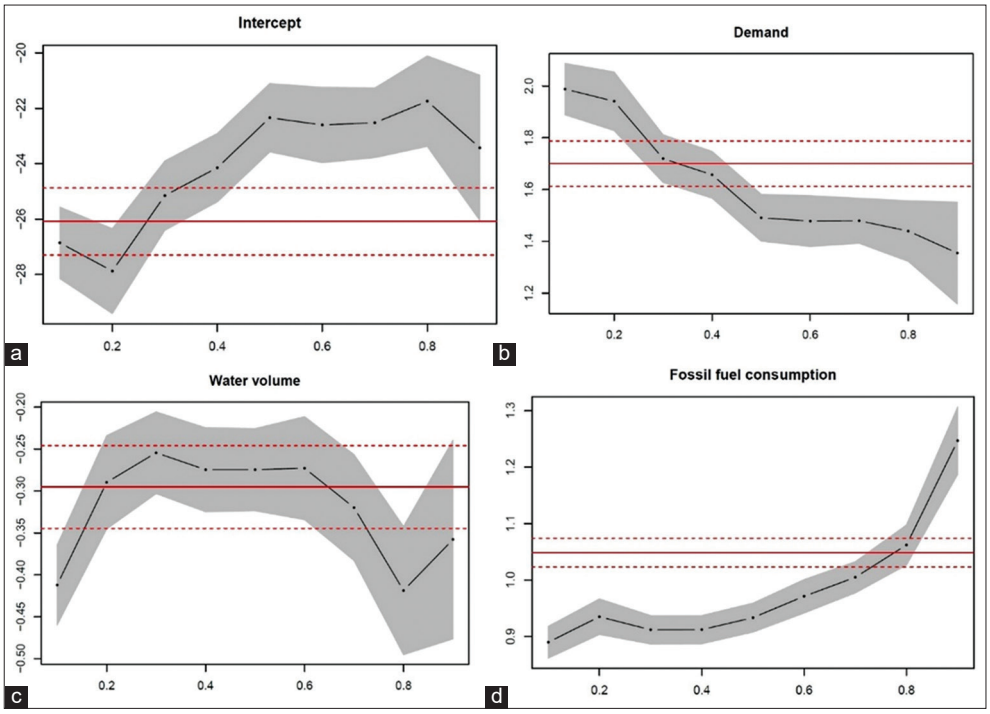
Source: authors' analysis.

Figure 6: Quantiles for electricity spot price for August 2009 – December 2019.



Source: authors' analysis.

Figure 7: Fundamental variables effects on the electricity spot price for different percentiles. the vertical axis in each subplot corresponds to the spot price response by effects of predictors, while The horizontal axis corresponds to quantiles, from the 10th to the 90th percentile. The dotted black lines represent the quantile regression coefficients and the gray area is the 95% confidence interval. The continuous red line is the linear regression coefficient estimated by OLS, and the discontinuous red lines are the 95% confidence intervals. The variables are defined as follows: (a) effects of the intercept; (b) effects of the demand; (c) effects of the water volume or reservoir capacity; (d) effects of the gas and coal consumption for generating



Source: Authors' analysis.

prices in the 60th and the 80th percentile, the demand tries to stabilize, but for prices over the 80th percentile, the effect associated with variation in demand is lower. Therefore, with a demand variation of 1%, the price variation is equal to 1.4%. Given the inverse relationship between prices and demand, its impact is less on high quantiles.

According to Barrientos et al. (2012), Barrientos-Marín and Toro-Martínez (2017), and García et al. (2011), demand is one of the most relevant spot price determinants. It is concluded that the price has a positive trend in the future by a positive demand shock. However, the effect is higher in the short-term. Besides, the price captures the complex effects of supply and demand

activity through the influence of the operational determinants: technological and organizational (Díaz-Contreras et al., 2014; Girish and Vijayalakshmi, 2013).

5.2.1.2. Water volume effects

The elasticity of the water volume is negative and significant statistically, independent of the quantiles. Therefore, an increased impact of water volume sensitivity was observed on lower and higher quantiles, i.e. when the water reservoir capacity is high, it always leads to a reduction in the electricity prices. In the first quantile, the price is low by a high water volume. It was observed that a water volume variation of 1% causes a price variation equal to −0.41%. In the last quantiles, the impact is higher because water volume becomes the most important source and an alternative to reduce the spot price when the thermal plants are on. In the 20th-70th percentiles the effects measured by quantile regression are similar to the median effects.

Hydraulic technology presents lower generation costs than thermal technology. However, hydric sources are high uncertainty to the energy and market reliability. Given the seasonal patterns in hydric sources, the electricity spot prices are lower in the rainy season and higher in the dry season (García et al., 2011). According to Barrientos-Marín and Toro-Martínez (2017) a positive effect on the available hydric capacity causes a negative real price. Likewise, hydropower generation depends on the future situation (or not observable); hence, this sector tries to influence on the spot prices.

5.2.1.3. Fossil fuel consumption effects

Positive and significant elasticities were observed for fossil fuel consumption. Around the 10th percentile, the effects are minor, but for prices over the 40th percentile, the effects are becoming higher. This means that the thermal plants must turn on by a decrease in water volume or an increase in demand and, as a result, the generation costs and spot prices increase. In the 90th percentile, the price variation is approximate 1.25% when the fossil fuel consumption is 1%.

According to Mosquera-López et al. (2017a), when the thermal generation plants are on, they present marginal costs of up to 300%, higher than hydropower plants. Therefore, the marginal generation costs show a relevant difference between the two most important generation technologies, which explains the price fluctuations.

6. CONCLUSIONS

Considering the Colombian power generation market structure, where hydropower generation is the most relevant source, followed by thermal power technology, a set of market fundamentals was validated through a price prediction using a machine learning trained model. Besides, by using quantile regression, the non-linear effects of these variables on the spot price were measured. In the sensitivity analyses for the different variables across the price distribution, it was observed how the demand, the water reservoir capacity, and the fossil fuel consumption influence the price.

Therefore, positive changes were observed in the spot price through demand variations. When the electricity consumption increases, all generation technologies must produce to meet demand. However, if the demand is not cover, the thermal power generation plants must turn on, affecting the price. By contrast, the elasticity of the water volume or reservoir capacity is negative, with increased impact on lower and higher quantiles. That is, seasonal patterns of reservoirs cause a strong price fluctuation, e.g., each rainy season, the spot price decrease significantly. An important aspect is the generation sector's influence on the price by future speculation of water volume; for this reason, it must be added a fundamental that captures the oligopoly structure.

Positive elasticities were found for fossil fuel consumption. It was revealed how gas and coal increased the price significantly on last quantiles. Exogenous effects such as dry seasons or the demand changes, increase the spot price through generation costs.

Therefore, it has described how the magnitude changes in fundamental variables in a hydrothermal power, explain the electricity spot price. The effect of reservoir changes represents the main risk factor for generators. Besides, the generation sector faces risk by fossil fuel price fluctuation; hence, they cannot recover the costs through the electricity price increases. Likewise, this study allowed identifying the importance of renewable energy because they can become a smoother of the volatility prices and prevent their extreme changes caused by exogenous effects.

Finally, to improve the model prediction it will be required the inclusion of generation concentration index or agent strategies. However, the model can serve as a point of reference, given the hydrothermal generation sector characteristic and exogenous factors that explain the electricity price dynamics.

REFERENCES

Aggarwal, S.K., Saini, L.M., Kumar, A. (2009), Electricity price forecasting in deregulated markets: A review and evaluation. International Journal of Electrical Power and Energy Systems, 31(1), 13-22.

Aye, S.A., Heyns, P.S. (2017), An integrated Gaussian process regression for prediction of remaining useful life of slow speed bearings based on acoustic emission. Mechanical Systems and Signal Processing, 84, 485-498.

Barria, C., Rudnick, H. (2011), Investment under uncertainty in power generation: Integrated electricity prices modeling and real options approach. IEEE Latin America Transactions, 9(5), 785-792.

Barrientos, J., Rodas, E., Velilla, E. (2012), Modelo para el pronóstico del precio de la energía eléctrica en Colombia. Lecturas de Economía, 77, 91-127.

Barrientos-Marín, J., Toro-Martínez, M. (2017), Análisis de los fundamentales del precio de la energía eléctrica: Evidencia empírica para Colombia. Revista de Economía del Caribe, 19, 34-63.

Berrie, T.W., Hoyle, M. (1985), Treating energy as a commodity. Energy Policy, 13(6), 506-510.

Blazsek, S., Hernández, H. (2018), Analysis of electricity prices for Central American countries using dynamic conditional score models. Empirical Economics, 55(4), 1807-1848.

Botero-Duque, J.P., García, J.J., Velásquez, H. (2016), Efectos del cargo por confiabilidad sobre el precio spot de la energía eléctrica en Colombia. Cuadernos de Economía, 35(68), 491-519.

Castaño, E., Sierra, J. (2012), Sobre la existencia de una raíz unitaria en la serie de tiempo mensual del precio de la electricidad en Colombia. Lecturas de Economía, 76, 259-291.

Castelli, M., Groznik, A., Popovič, A. (2020), Forecasting electricity prices: A machine learning approach. Algorithms, 13(5), 119.

Ciarreta, A., Lagullón, M., Zarraga, A. (2011), Modelación de los precios en el mercado eléctrico español. Cuadernos de Economía, 30, 227-250.

Cotia, B.P., Borges, C.L.T., Diniz, A.L. (2019), Optimization of wind power generation to minimize operation costs in the daily scheduling of hydrothermal systems. International Journal of Electrical Power and Energy Systems, 113, 539-548.

Crespo-Cuaresma, J., Hlouskova, J., Kossmeier, S., and Obersteiner, M. (2004), Forecasting electricity spot-prices using linear univariate time-series models. Applied Energy, 77(1), 87-106.

Deng, S., Oren, S. (2006), Electricity derivatives and risk management. Energy, 31(6-7), 940-953.

Díaz, G., Coto, J., Gómez-Aleixandre, J. (2019), Prediction and explanation of the formation of the Spanish day-ahead electricity price through machine learning regression. Applied Energy, 239, 610-625.

Díaz-Contreras, J.A., Macías-Villalba, G.I., Luna-González, E. (2014), Estrategia de cobertura con productos derivados para el mercado energético colombiano. Estudios Gerenciales, 30(130), 55-64.

Fernández-Blanco, R., Kavvadias, K., Hidalgo González, I. (2017), Quantifying the water-power linkage on hydrothermal power systems: A Greek case study. Applied Energy, 203, 240-253.

Gao, C., Bompard, E., Napoli, R., Wan, Q., Zhou, J. (2008), Bidding strategy with forecast technology based on support vector machine in the electricity market. Physica A: Statistical Mechanics and Its Applications, 387(15), 3874-3881.

García, J., Gaviria, A., Salazar, L. (2011), Determinantes del precio de la energía eléctrica en el mercado no regulado en Colombia. Ciencias Estratégicas, 19, 225-246.

Girish, G.P., Vijayalakshmi, S. (2013), Determinants of electricity price in competitive power market. International Journal of Business and Management, 8(21), 70-75.

Gonzalez-Briones, A., Hernandez, G., Corchado, J.M., Omatu, S., Mohamad, M.S. (2019), Machine Learning models for electricity consumption forecasting: A review. In: 2019 2nd International Conference on Computer Applications and Information Security (ICCAIS). p1-6.

He, Y.X., Zhang, S.L., Yang, L.Y., Wang, Y.J., Wang, J. (2010), Economic analysis of coal price-electricity price adjustment in China based on the CGE model. Energy Policy, 38(11), 6629-6637.

Huisman, R., Mahieu, R. (2003), Regime jumps in electricity prices. Energy Economics, 10, 425-434.

Imani, M.H., Bompard, E., Colella, P., Huang, T. (2020), Predictive methods of electricity price: An application to the Italian electricity market. In: 2020 IEEE International Conference on Environment and Electrical Engineering and 2020 IEEE Industrial and Commercial Power Systems Europe (EEEIC/I and CPS Europe). p1-6.

Kian, A., Keyhani, A. (2001), Stochastic price modeling of electricity in deregulated energy markets. In: Proceedings of the 34th Annual Hawaii International Conference on System Sciences. p7.

Koenker, R. (2004), Quantile regression for longitudinal data. Journal of Multivariate Analysis, 91(1), 74-89.

Koenker, R., Bassett, G. (1978), Regression quantiles. Econometrica, 46(1), 33.

Lira, F., Muñoz, C., Núñez, F., Cipriano, A. (2009), Short-term forecasting of electricity prices in the Colombian electricity market. IET Generation, Transmission and Distribution, 3(11), 980-986.

Ma, L., Koenker, R. (2006), Quantile regression methods for recursive structural equation models. Journal of Econometrics, 134(2), 471-506.

Maciejowska, K. (2020), Assessing the impact of renewable energy sources on the electricity price level and variability-a quantile regression approach. Energy Economics, 85, 104532.

Mandal, P., Senjyu, T., Urasaki, N., Funabashi, T., Srivastava, A.K. (2007), A novel approach to forecast electricity price for PJM using neural network and similar days method. IEEE Transactions on Power Systems, 22(4), 2058-2065.

Montes, C. (2018), La incertidumbre climática y el dilema energético colombiano. Revista de la Academia Colombiana de Ciencias Exactas, Físicas y Naturales, 42(165), 392-401.

Mosquera-López, S., Manotas-Duque, D.F., Uribe, J.M. (2017a), Risk asymmetries in hydrothermal power generation markets. Electric Power Systems Research, 147, 154-164.

Mosquera-López, S., Nursimulu, A. (2019), Drivers of electricity price dynamics: Comparative analysis of spot and futures markets. Energy Policy, 126, 76-87.

Mosquera-López, S., Uribe, J.M., Manotas-Duque, D.F. (2017b), Nonlinear empirical pricing in electricity markets using fundamental weather factors. Energy, 139, 594-605.

Quintero-Quintero, M.C., Isaza-Cuervo, F. (2013), Dependencia hidrológica y regulatoria en la formación de precio de la energía en un sistema hidrodominado: Caso sistema eléctrico colombiano. Revista Ingenierías Universidad de Medellín, 12(22), 85-95.

Rasmussen, C.E., Williams, C.K.I. (2006), Gaussian Processes for Machine Learning. United States: MIT Press.

Ribeiro, M., Stefenon, S., de Lima, J., Nied, A., Mariani, V., Coelho, L. (2020), Electricity price forecasting based on self-adaptive decomposition and heterogeneous ensemble learning. Energies, 13(19), 5190.

Rodriguez, C.P., Anders, G.J. (2004), Energy price forecasting in the ontario competitive power system market. IEEE Transactions on Power Systems, 19(1), 366-374.

Samudio-Carter, C., Vargas, A., Albarracín-Sánchez, R., Lin, J. (2019), Mitigation of price spike in unit commitment: A probabilistic approach. Energy Economics, 80, 1041-1049.

The Mathworks, Inc. (2020), Statistics and Machine Learning Toolbox User's Guide. United States: The Mathworks, Inc. Available from: https://www.la.mathworks.com/help/pdf_doc/stats/stats.pdf.

Uribe, J.M., Guillen, M. (2020), Quantile Regression for Cross-Sectional and Time Series Data: Applications in Energy Markets Using R. Berlin: Springer International Publishing.

Vaca, J., Núñez, G., Kido, A. (2019), Análisis multisectorial del incremento de precios de la electricidad en la economía de México. Problemas del Desarrollo. Revista Latinoamericana de Economía, 50(196), 167-189.

Vidal, P., Sierra, L., Cerón, J. (2020), Demanda Nacional de Energía y Crecimiento Económico en Tiempos de Cuarentena. Colombia: Pontificia Univerdiad Javeriana.

Weron, R. (2014), Electricity price forecasting: A review of the state-of-the-art with a look into the future. International Journal of Forecasting, 30(4), 1030-1081.

Xavier, E.M., Pereira, G.M., Friedrich, L.R., Schneider, L.C., Danesi, L.C., Borchardt, M. (2016), Requirements to leverage the electricity distributors' sales and revenues in the brazilian free market. IEEE Latin America Transactions, 14(10), 4293-4303.

Zhang, L., Luh, P.B. (2005), Neural network-based market clearing price prediction and confidence interval estimation with an improved extended kalman filter method. IEEE Transactions on Power Systems, 20(1), 59-66.

Zhang, Y., Zhou, Q., Sun, C., Lei, S., Liu, Y., Song, Y. (2008), RBF neural network and ANFIS-based short-term load forecasting approach in real-time price environment. IEEE Transactions on Power Systems, 23(3), 853-858.

Ziel, F. (2016), Forecasting electricity spot prices using lasso: On capturing the autoregressive intraday structure. IEEE Transactions on Power Systems, 31, 4977-4987.

APPENDIX A

Table A.I shows the quantile regression coefficients from 10th to 90th percentiles. All coefficients are significant statistically for a 1% level of confidence.

Table A.I: Quantile regression coefficients for different quantiles

Predictors	$\beta^{0.1}$	$\beta^{0.2}$	$\beta^{0.3}$	$\beta^{0.4}$	$\beta^{0.5}$	$\beta^{0.6}$	$\beta^{0.7}$	$\beta^{0.8}$	$\beta^{0.9}$
Intercept	−26.857	−27.877	−25.152	−24.142	−22.338	−22.600	−22.521	−21.739	−23.425
Demand	1.988	1.941	1.719	1.657	1.491	1.478	1.479	1.439	1.355
Water volume	−0.412	−0.289	−0.254	−0.275	−0.275	−0.273	−0.319	−0.419	−0.358
Fossil fuel consumption	0.899	0.935	0.912	0.912	0.933	0.971	1.005	1.062	1.247

Source: authors' analysis