Paper accepted for presentation at 2009 IEEE Bucharest Power Tech Conference, June 28th - July 2nd, Bucharest, Romania

Pricing Analysis in the Brazilian Energy Market: a Decision Tree Approach

J. C. Reston Filho, C. M. Affonso e R. C. L. Oliveira

***Abstract*-- There is a general consensus that electricity price forecasting is an important task nowadays since power market players are interested in the maximization of profit and the minimization of risk. In this context, this paper proposes the use of Data-Mining techniques to predict the short-term electricity price in the Brazilian market. In Brazil, the market model adopted has unique characteristics with a centralized form of dispatch due to the predominance of hydro generation. To apply the proposed prediction model, all features of the Brazilian electricity market are considered, such as the transmission restrictions among geo-electrical regions and the price dependency with storage energy in reservoirs. In the proposed prediction model, the electricity price is the dependent variable and the monthly time series data sets from the Brazilian system (such as power load, stored energy and thermal generation) are the independent variables. First, clustering of the data samples is performed to group similar behavior of the attributes. After that, a decision tree algorithm is applied to extract if-then rules from database. The rules obtained allow the identification of attributes that most influence the short-term electricity price. Results show that the proposed model can be an attractive tool to all electricity market players to forecast the short-term electricity price and mitigate the risks in purchasing power.**

***Index Terms*—Clustering, Data-Mining, Decision Trees, Short-Term Electricity Price Forecasting.**

I. INTROD\UCTION

IN recent years, the electric power industry deregulation and free competition has been widely spread in the world. In this new model, the understanding of electric power supply as a public service is being replaced by the notion of a competitive market. In this context, market players are interested in the maximization of profit and the minimization of risk.

The electricity price exhibit high volatility since is influenced by many factors such as the balance between the demand and the available supply for electricity, the economic growth, the weather, the power-plant mix, the prices of fuels and others [1].

For this reason, generators face price risks because they sell energy at variable pool prices while their fuel and other costs are fixed. If a producer has an accurate forecast of the prices, it can develop a bidding strategy to maximize its profit. Distributors also face price risks because they supply energy to most of their costumers at an annual fixed tariff but they have to purchase electricity at variable pool price. Then, distributors can make a plan to minimize his electricity cost if an accurate price forecast is available. In the medium and long-term period, generators and distributors also need price forecasting for the decisions on selling or buying energy in the pool or through bilateral contracts. Facing this reality, electricity price forecast is therefore extremely important for all electricity markets players, namely for risk management.

Many researches have been done for electricity-price forecasting. In those works, we can find a variety of techniques based on the analysis algorithm of stochastic time series, such as ARIMA (autoregressive integrated moving average) methodology, and intelligent algorithms such as artificial neural networks, support-vector machine and data mining [2,3,4,5]. However, price-forecasting techniques are still in their early stages of maturity [6].

This paper proposes the use of Data-mining techniques based on clustering and decision trees to estimate the short-term electricity price in the Brazilian market. The decision tree is useful for extracting *if-then* rules from database. The main advantage of decision trees is the easy interpretability of the results [7]. The results show that this model can be a competitive advantage for the participants in the market to predict the short term electricity price. The main contribution of this paper is the application of Data-mining techniques to forecast the short-term electricity price of the Brazilian system, which has unique characteristics and was never investigated before.

This paper is organized as follows. Section II describes the test system used to study the price-forecasting problem. Section III presents the methodology used in this paper and practical details of the data pre-processing, clustering and data mining model. The results are presented and discussed in Section IV. Finally, the main conclusions are summarized in Section V.

II. BRAZILIAN ENERGY SYSTEM

The proposed prediction model was tested in the Brazilian electricity market. The Brazilian National Interconnected System (SIN) has an installed capacity of 88GW, from which hydropower responds for more than 84.31%, as shows Fig. 1. These data are referred to the operation of December 2007.

The hydro system is composed of several large reservoirs

This work was supported in part by FAPEAM – AM, Brazil.

J. C. Reston Filho is with Instituto Dados da Amazônia – IDAAM, Manaus, AM, 69050-010, Brazil (e-mail: jcreston@gmail.com)

C. M. Affonso and R. C. L. Oliveira are with the Faculty of Electrical Engineering, Federal University of Para, Belem, PA, 66075-110 Brazil (e-mail: carolina@ufpa.br, limao@ufpa.br).

978-1-4244-2235-7/09/$25.00 ©2009 IEEE

capable of multi-year regulation, organized in a complex topology over several basins. Thermal generation includes natural gas, coal and diesel plants. The country is fully interconnected at the bulk power level by a 80,000 km meshed high-voltage transmission network, with voltages ranging from 230 kV to 765 kV ac, plus two 600 kV dc links connecting the binational Itaipu power plant (14,000 MW) to the main grid.

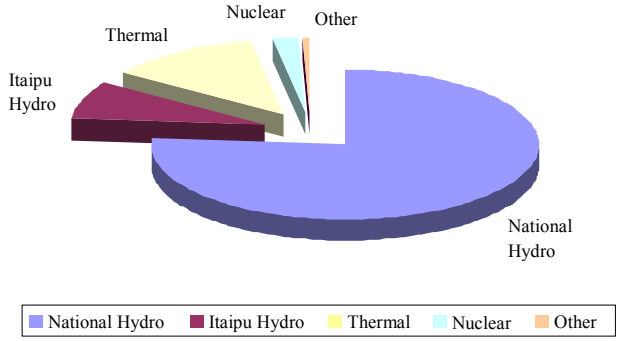


Fig. 1. Installed capacity of Brazilian electrical system.

The National System Operator (ONS) is responsible for the proper operation and control of the Brazilian generation and transmission system. The electricity transactions (short term market and bilateral contracts) among the companies in the Brazilian system are conducted by the Electrical Energy Commercialization Chamber (CCEE).

Due to the predominance of the hydro generation, the model adopted is a centralized form of dispatch called tight pool. The transmission restrictions among geo-electrical regions of the SIN impose the division of the energy market into four submarkets: North, Northeast, Center-east/Southeast and South. These markets can import/export energy from/to each other. Fig. 2 shows the Brazilian electricity market configuration.

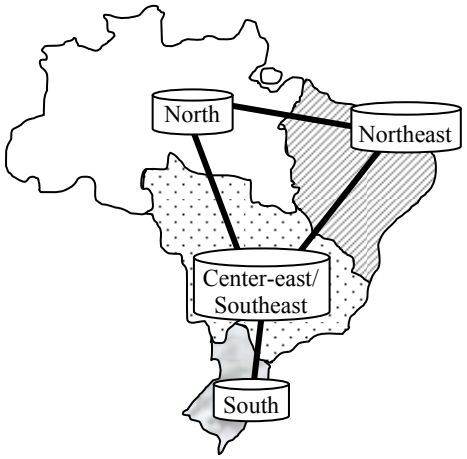


Fig. 2. Brazilian electricity market configuration.

The CCEE makes the accounting of the differences between actual production/consumption and contracts. Positive or negative differences are settled in the spot market through the spot price, which is named PLD (Settlement Price for the Differences). Fig. 3 illustrates this commercialization process. The PLD price is set weekly, for each load level in each submarket. Its basis is the marginal cost to operate the system, within a given interval.

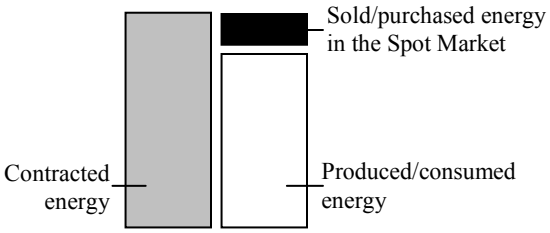


Fig. 3. Commercialization process in the Spot Market.

The PLD is evaluated by using a hydro-thermal dispatch. The mathematical model minimizes the marginal price of the system operation considering the future cost. Although the use of the water in the operation decreases the immediate operation cost, it would increase the future operation cost.

The spot price PLD depends on the storage in reservoirs. Because storage depends on inflows, which is a random variable, there is a great volatility on the Brazilian spot price. In order to illustrate the relationship between the PLD, storage and inflow energy, Fig. 4 shows the behavior of these variables for a period from May 2006 through April 2007 to the Center-east/Southeast region.

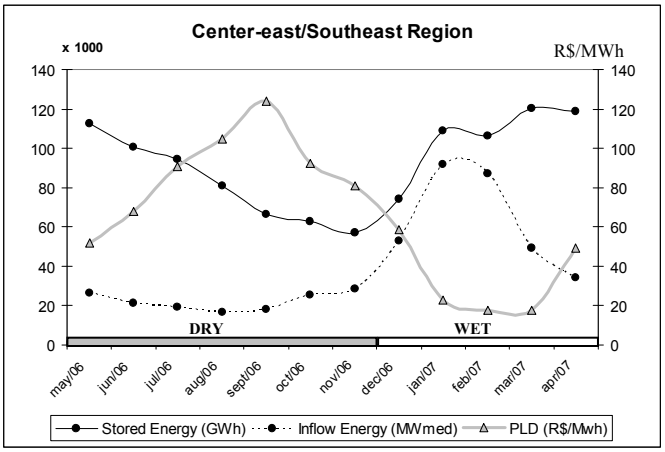


Fig. 4. Relationship between PLD, stored and inflow energy.

In this region the dry season varies from May through November and the wet season varies from December through April. The inflow energy is lower during the dry season and it increases during the wet season. This behavior is also seen in stored energy. However, note that there is a delayed relationship between inflow and storage energy. Also, during the dry season, the PLD tends to be higher due to the use of the thermal power plants.

III. Methodology

The process of extracting knowledge from data consists of several steps in order to obtain consistent patterns and/or systematic relationships between variables, from a large database. The process consists of four basic stages as shows Fig. 5: data selection, data pre-processing, Data Mining (DM) and interpretation/evaluation of the discovery knowledge.

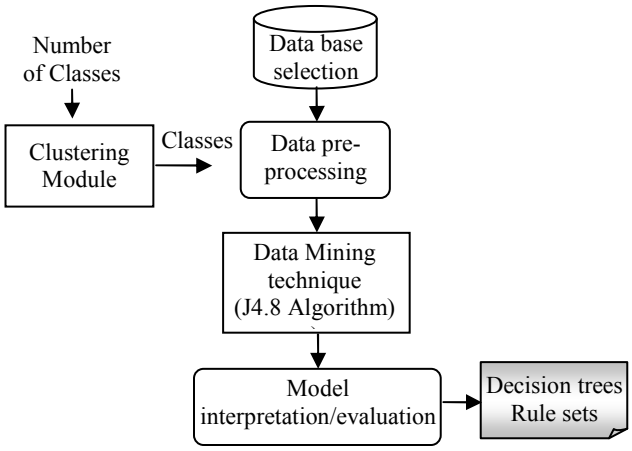


Fig. 5. Extracting knowledge process.

In the first stage data with more significance to the process must be selected. A data set of significant dimension is needed.

The second stage is the preprocessing. The data set is first cleaned. It is required in order to detect and correct bad data as well as remove outliers. Therefore, the process of extracting knowledge will not be influenced by atypical data. In the clustering module, clustering techniques are used to obtain a number of classes that group similar behavior of the attributes. The main goal of clustering data is to find common patterns or to group similar cases in the data.

Once the available data is collected and pre-processed, data mining techniques (processing algorithms) are applied to the data set. This is made using one isolated technique or combining several techniques, to build a model able to find relevant knowledge found in the data.

Finally, once the model is obtained the interpretation of the discovered knowledge is improved. Also, the obtained model is evaluated applying the test data set. This knowledge can provide new insights into relationships between data elements and facilitate more productive and sophisticated decision support applications.

The following sections describe in detail each step of the methodology using a realistic case study.

A. Data Base Selection

This research used 11 variables to create the database. These variables are indicated in Table I. The historical data are available in Brazilian Electrical Energy Commercialization Chamber website [8] and in National System Operator website [9]. The PLD is the goal attribute and the others variables are the input attributes.

Monthly samples were recorded between May 2003 and December 2008. A database containing these variables for this period was constructed for each one of the four submarkets (North, Northeast, South and Center-east/Southeast).

First, the attributes Itaipu hydro generation (ItHyGen), total system hydro generation (TotHyGen) and total system thermal generation (TotTherGen) were not used to create the model. However, the results showed that these are important attributes that should appear in the database, with high correlation with PLD, since these data give the implicit information of the energy transferred between the submarkets.

TABLE I
ATTRIBUTES DESCRIPTION

Attribute	Definition	Unit
PLD	Short-term energy price of the submarket	R$/MWh
StEn	Stored energy of the submarket	GWh
InEn	Inflow energy of the submarket	MWmed
HyGen	Hydro generation of the submarket	MWmed
ItHyGen	Itaipu hydro generation	MWmed
TotHyGen	Total system hydro generation	MWmed
TherGen	Thermal generation of the submarket	MWmed
TotTherGen	Total system thermal generation	MWmed
Load	Power load of the submarket	MWmed
Seas	Season	wet/dry
Mon	Month	Jan – Dec

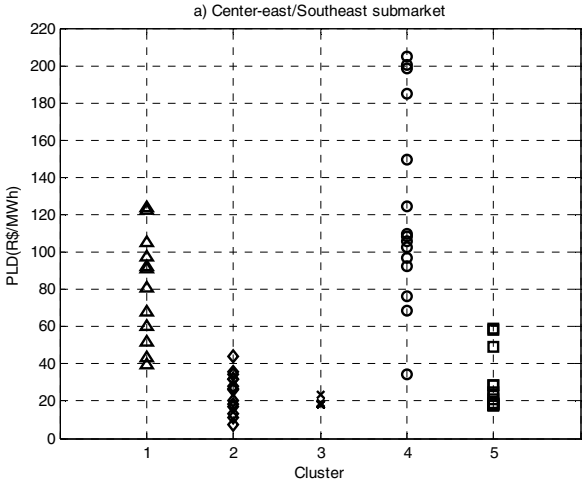
B. Data Pre-processing

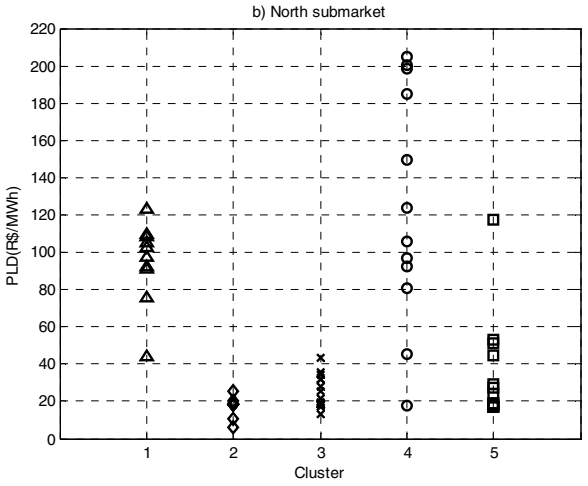
In the pre-processing, the noises were eliminated mainly due to periods of energy rationing in Brazil. We also filtered the records with missing data. Then, the data were stored in 67 instances for each submarket, totaling 4 sets of data. Each instance contains all attributes described in Table I.

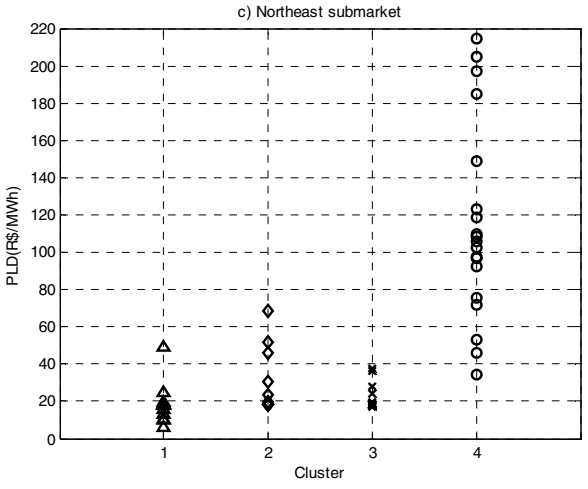
This paper used the software WEKA (Waikato Environment for Knowledge Analysis – WEKA), which is a widely used data mining tool with several algorithms. Several established clustering algorithms [10] available in WEKA [11] were compared to find the best number of clusters. The algorithm with best performance was the Expectation-maximization algorithm (EM Algorithm) [5]. In the expectation step, EM algorithm estimates the cluster to which each instance belongs given the distribution parameters. In the second step, EM estimates the parameters from the classified instances [10]. In the EM algorithm instances are assigned to classes probabilistically rather than categorically. Since the number of clusters is not known in advance, we therefore tried different possibilities. Table II presents the number of clusters for each submarket and also the mean, standard deviation and number of samples of each cluster. Fig. 6 presents the PLD grouped for each cluster for each one of the submarkets.

TABLE II
CHARACTERISTICS OF THE CLUSTERING MODULE

Submarket		Cluster1	Cluster2	Cluster3	Cluster4	Cluster5
North	Mean	94.94	17.57	23.48	125.09	35.00
	StdDev	21.91	4.83	7.95	63.02	28.12
	Samples	10	12	20	12	13
Northeast	Mean	18.65	27.15	21.17	115.15	
	StdDev	10.07	15.68	6.33	53.17	---
	Samples	14	15	18	19	
South	Mean	99.78	59.10	59.90	20.44	
	StdDev	64.51	38.72	33.16	5.98	---
	Samples	19	6	18	24	
Center-east/ Southeast	Mean	81.24	24.66	19.03	123.91	28.65
	StdDev	29.03	10.55	1.47	52.51	15.09
	Samples	12	14	12	15	14







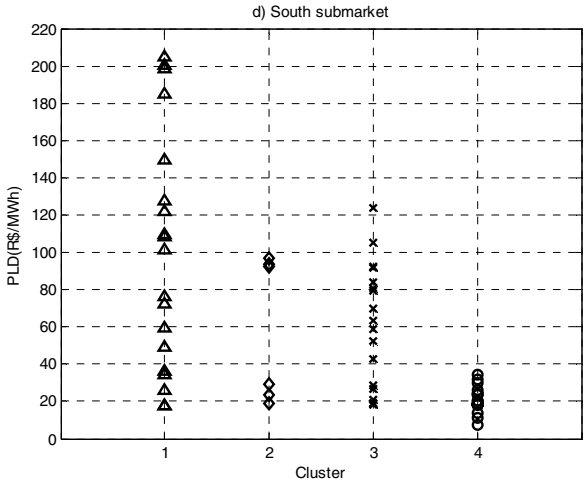


Fig. 6. PLD for each cluster for a) Center-east/Southeast submarket, b) North submarket, c) Northeast submarket and d) South submarket.

C. Data Mining Technique

The algorithm used was the J4.8 available in WEKA [10]. This algorithm is a supervised approach to the classification problem which generates a decision tree. The decision trees are easy to understand and easily converted to a set of production rules. They can classify both categorical and numerical data, but the goal attribute must be categorical.

D. Model Interpretation/Evaluation

The model evaluation is performed using the 10-fold Cross-Validation strategy [12]. This type of evaluation was selected to improve the results obtained in the presence of small data sets. Using this evaluation technique it is possible to train the algorithm using the entire data set and obtain a more precise model. This will increase the computational effort but improves the model's capacity of generalization to different data sets. The evaluation is performed by randomly splitting the initial sample in ten subsamples. The model is trained using 90% of the data set and tested with the 10% left. This process (train and test) is performed 10 times on different training sets. The cross validation refers to the use of different test sets at each stage, so that at the end of their performance all the data have been used.

IV. Results Analysis

The methodology presented in this paper was applied to data from the Brazilian Electricity Market. Four decision trees were obtained, one for each submarket. We will first present, with more details, the results for the Center-east/Southeast submarket.

Fig. 7 shows the decision tree obtained for the Center-east/Southeast submarket. The non-terminal nodes represent test/decisions on one or more attributes and the terminal nodes reflect decision outcomes. Table III shows the rule set that summarizes this decision tree.

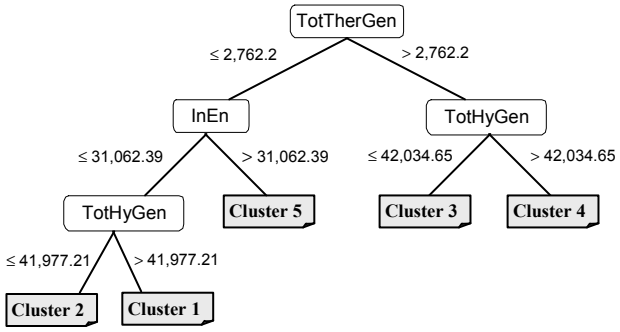


Fig. 7. Decision tree for the Center-east/Southeast submarket.

TABLE III
RULE SET FOR THE CENTER-EAST/SOUTHEAST SUBMARKET

1. If TotTherGen ≤ 2,762.2	**and InEn** ≤ 31,062.39	**and TotHyGen** >41,977.21	**Then** PLD: Cluster 1
2. If TotTherGen ≤ 2,762.2	**and InEn** ≤ 31,062.39	**and TotHyGen** ≤41,977.21	**Then** PLD: Cluster 2
3. If TotTherGen > 2,762.2		**and TotHyGen** ≤42,034.65	**Then** PLD: Cluster 3
4. If TotTherGen > 2,762.2		**and TotHyGen** >42,034.65	**Then** PLD: Cluster 4
5. If TotTherGen ≤ 2,762.2	**and InEn** > 31,062.39		**Then** PLD: Cluster 5

Table IV presents the confusion matrix obtained for this submarket. A confusion matrix [5] contains information about actual and predicted classifications done by a classification system. The performance of such systems is commonly evaluated using the data in the matrix.

The confusion matrix shows that 12 instances from Cluster 1, 12 instances from Cluster 2, 11 instances from Cluster 3, 15 instances from Cluster 4 and 13 instances from Cluster 5 were correctly classified. Only 2 records from Cluster 2, 1 record from Cluster 3 and 1 record from Cluster 5 were misclassified. Then, 63 instances were correctly classified and 4 instances were misclassified. The last column in Table III shows the rate of cases correctly classified to each class.

TABLE IV
CONFUSION MATRIX FOR THE CENTER-EAST/SOUTHEAST SUBMARKET

Actual	Classified					
	Cluster1	**Cluster2**	**Cluster3**	**Cluster4**	**Cluster5**	**Correct Rate**
Cluster1	12	0	0	0	0	1.000
Cluster2	0	12	0	1	1	0.857
Cluster3	0	1	11	0	0	0.917
Cluster4	0	0	0	15	0	1.000
Cluster5	0	0	0	1	13	0.929

Tables V, VI and VII shows the rule set obtained with the model for submarkets North, Northeast and Southeast respectively. The rule set obtained are quite simple to understand for all submarkets.

TABLE V
RULE SET FOR THE NORTH SUBMARKET

1. If Load > 3,215.39	**and StEn** > 4,575.6	**and InEn** ≤ 4,359.0		**Then** PLD: Cluster 1
2. If Load ≤ 3,215.39			**and HyGen** > 3,624.09	**Then** PLD: Cluster 2
3. If Load ≤3,215.39			**and HyGen** ≤ 3,624.09	**Then** PLD: Cluster 3
4. If Load > 3,215.39	**and StEn** ≤ 4,575.6			**Then** PLD: Cluster 4
5. If Load > 3,215.39	**and StEn** > 4,575.6	**and InEn** > 4,359.0		**Then** PLD: Cluster 5

TABLE VI
RULE SET FOR THE NORTHEAST SUBMARKET

1. If Load ≤ 6,388.42				**Then** PLD: Cluster 1
2. If Load > 6,388.42	**and InEn** ≤ 5.366.32	**and TotHyGen** ≤43,740.76		**Then** PLD: Cluster 2
3. If Load > 6,388.42	**and InEn** > 5.366.32		**and TotTherGen** ≤ 3,287.03	**Then** PLD: Cluster 3
4. If Load > 6,388.42	**and InEn** ≤ 5.366.32	**and TotHyGen** >43,740.76		**Then** PLD: Cluster 4
5. If Load > 6,388.42	**and InEn** > 5.366.32		**and TotTherGen** > 3,287.03	**Then** PLD: Cluster 4

TABLE VII
RULE SET FOR THE SOUTH SUBMARKET

1. if TotHyGen > 41,977.21	**and StEn** ≤ 11,424.0	**and InEn** > 5,518.65		**Then** PLD: Cluster 1
2. if TotHyGen > 41,977.21	**and StEn** ≤ 11,424.0	**and InEn** ≤ 5,518.65	**and HyGen** > 5,770.59	**Then** PLD: Cluster 1
3. if TotHyGen > 41,977.21	**and StEn** > 11,424.0			**Then** PLD: Cluster 2
4. if TotHyGen > 41,977.21	**and StEn** ≤ 11,424.0	**and InEn** ≤ 5,518.65	**and HyGen** ≤ 5,770.59	**Then** PLD: Cluster 3
5. if TotHyGen ≤ 41,977.21				**Then** PLD: Cluster 4

Table VIII presents the overall accuracy obtained for each one of the four submarkets with the algorithm J4.8. The overall accuracy is evaluated dividing the sum of the correct classifications (diagonal elements of the confusion matrix) by the total number of instances.

The characteristics of the classification models obtained for each submarket are also presented in Table VIII. The model, according to the data set, selected different relevant attributes to create the decision trees. It is possible to conclude that total system hydro generation, total system thermal generation, power load, hydro generation, stored energy and inflow energy of each submarket are the most relevant attributes to predict the short-term electricity price. Moreover, total system hydro and thermal generation are significant attributes to determine the short-term energy price for all regions, except for the North submarket. This can be explained since this region has a huge capacity of generation, being less dependent on importation of energy to meet its own demand.

TABLE VIII
CHARACTERISTICS OF THE CLASSIFICATION MODELS

Submarket	Overall Accuracy	Relevant Attributes	Number of Rules
North	97.01%	Load, HyGen, StEn, InEn	5
Northeast	87.87%	Load, InEn, TotHyGen, TotTherGen	5
South	89.55%	TotHyGen, HyGen, StEn, InEn	5
Center-east/ Southeast	94.02%	TotHyGen, TotTherGen, InEn	5

V. CONCLUSIONS

This paper presented the application of a data-mining algorithm in the Brazilian electrical system. The proposed model is based on clustering technique and decision trees to estimate the short-term electricity price. The Brazilian system, which has a centralized form of dispatch with unique characteristics, was carefully examined to apply the prediction model.

The rules obtained are quite simple to understand and show that the most important attributes to the classification of the short-term electricity price are: load, stored energy, inflow energy, hydro and thermal generation of the submarket and total system hydro and thermal generation. Also, the energy transferred between the submarkets appeared to be relevant to determine the energy price.

The results show that the suggested approach allows the classification and prediction of the short-term electricity price with high correct rate. The decision tree offers an attractive alternative to forecast and to explain the short-term energy prices behavior. The proposed model can be an attractive tool to all electricity market players to forecast the short-term electricity price and mitigate the risks in purchasing power.

VI. REFERENCES

[1] M. Shahidehpour, and M. Alomoush, "Restructured Electrical Power Systems – Operation, Trading, and Volatility", Marcel Dekker, Inc, New York, 2001.

[2] J. Contreras, R. Espinola, F.J. Nogales, and A.J. Conejo, "ARIMA models to predict next-day electricity prices", *IEEE Trans. Power Syst*., vol. 18 (3), pp. 1014–1020, 2003.

[3] C.P. Rodriguez, and G.J. Anders, "Energy prices forecasting in the Ontario competitive power system market", *IEEE Trans. Power Syst*., vol. 19 (1), pp. 366–374, 2004.

[4] D. Zhou, F. Gao, and X. Guan, "Application of accurate online support vector regression in energy price forecast", in *Proc. 2004 Fifth World Congress an Intelligent Control and Automation*, pp. 1838–1842.

[5] X. Lu, Z. Yang and A. Dong, "Electricity market price spike forecast with data mining techniques", Electric power systems research, vol. 73, pp. 19-29, 2005.

[6] S. K. Aggarwal, L. M. Saini, A. Kumar, "Electricity price forecasting in deregulated markets: A review and evaluation", *Electrical Power and Energy Systems,* vol. 31, pp. 13-22, 2009.

[7] A. H. Nizar, Z. Y. Dong, and J. H. Zhao, "Load profiling and data mining techniques in electricity deregulated market," presented at Power Engineering Society General Meeting, 2006. IEEE, 2006.

[8] Brazilian Electrical Energy Commercialization Chamber (CCEE). "Valores Médios do PLD por submercado", Brasília, Brazil. Available: www.ccee.gov.br

[9] National System Operator (ONS), "Relatórios anuais" Rio de Janeiro, Brazil. Avaiable: www.ons.gov.br

[10] F. Martínez-Álvarez, A. Troncoso, J. C. Riquelme and J. M. Riquelme, "Partitioning-Clustering Techniques Applied to the Electricity Price Time Series", Intelligent Data Engineering and Automated Learning - IDEAL 2007, Vol. 4881, pp. 990-999, 2007.

[11] I. H. Witten and E. Frank, "Data mining: practical machine learning tools and techniques", 2nd ed. ed. San Francisco, Calif.: Morgan Kaufman, 2005.

[12] Hsiao-Tien Pao, "A Neural Network Approach to m-Daily-Ahead Electricity Price Prediction", Advances in Neural Networks - ISNN 2006, vol. 3972, pp. 1284-1289, 2006.

VII. BIOGRAPHIES

José Carlos Reston Filho received the B.S. degrees in electrical engineering from Universidade Federal do Amazonas, Brazil, in 1998, the Master's degree from Universidade Federal do Para, Brazil, in 2007. Currently, he is a Professor of Engineering in the Universidade Gama Filho, Brazil. His research interest is on intelligent data analysis applied to electricity market.

Carolina de Mattos Affonso received the B.S. degrees in electrical engineering from Universidade Federal do Para, Brazil, in 1998, the Master's degree from Universidade Federal de Santa Catarina, Brazil, in 2000 and the Ph.D. degree in electrical engineering from Universidade Estadual de Campinas (UNICAMP), Brazil, in 2004. Currently, she is a Professor of Electrical Engineering with the Universidade Federal do Para, Brazil. Her research interests are on power system stability analysis, power quality, distributed generation and electricity market.

Roberto Célio Limão de Oliveira received the B.S. degree in electrical engineering from Universidade Federal do Para, Brazil in 1987, the Master's degree in Electronic Engineering from Instituto Tecnologico de Aeronautica, Brazil, in 1991 and the Ph.D. degree in electrical engineering from Universidade Federal de Santa Catarina, Brazil, in 1999. Currently, he is Professor of Computer Engineering with the Universidade Federal do Para, Brazil. His research interests are on intelligent control and natural computing. He is the Director of Science and Technology of the State of Para, Brazil.