Forecasting Electric Load Demand through Advanced Statistical Techniques

Forecasting Electric Load Demand through Advanced Statistical Techniques

Jesús Silva[1], Alexa Senior Naveda[2], Jesús García Guliany[3], William Niebles Núñez[4], Hugo Hernández Palma[5]

[1]Universidad Peruana de Ciencias Aplicadas, Lima, Perú.
[2] Universidad de la Costa, Barranquilla, Atlántico, Colombia
[3]Universidad Simón Bolívar, Barranquilla, Atlántico, Colombia
[4]Universidad del Sucre, Sincelejo, Sucre, Colombia.
[5]Universidad del Atlántico, Puerto Colombia, Atlántico, Colombia.

[1]**Email**: jesussilvaUPC@gmail.com

Abstract. Traditional forecasting models have been widely used for decision-making in production, finance and energy. Such is the case of the ARIMA models, developed in the 1970s by George Box and Gwilym Jenkins [1], which incorporate characteristics of the past models of the same series, according to their autocorrelation. This work compares advanced statistical methods for determining the demand for electricity in Colombia, including the SARIMA, econometric and Bayesian methods.

1. Introduction

There are several studies on the habitual behavior of energy in Colombia [1]; however, shortcomings have been found in some of them, such as: lack of success, lack of integration with exogenous variables, lack of data to estimate models, among others.

In Colombia, the National Dispatch Center (NDC), department of XM Compañía de Expertos en Mercados S.A.E.S.P, a subsidiary of ISA, is in charge of the operation and administration of the entire National Interconnected System of Colombia (SIN) [2]. In other words, its task is to plan, monitor and control the resources of energy generators, transmitters, distributors and traders. The CND must make a maneuver plan for the generating companies, indicating the amount of power they must produce daily. For this reason, the forecast of energy demand is one of the most important tools in this process. That is why an effective prediction is really necessary, guaranteeing quality, security and reliability in the service for the users. In this regard, [3] states: "The prediction of demand is a problem of great importance for the electricity sector, since, based on its results, the agents of the energy market make the most appropriate decisions for their work". For its part, Codensa S.A. ESP, the electricity distribution and marketing company in Colombia, has prepared forecasts based on linear regression, exponential smoothing and moving average [4], comparing them using the ASM criterion [5].

This paper presents a characterization, analysis and comparison of the following models: SARIMA, econometric [6], and a Bayesian technique named Gaussian regression with Monte Carlo simulation by Markov Chains (MCMC) [7], which allow, after the estimation of tests, the validation and measurement of the average absolute percentage error indicator (MAPE) with adjustment and prognosis data for each one, and determine the best one to make the prediction of the demand for the Colombian state.

2. Method

The process started from an exploratory analysis of the time series described as a set of historical data of daily energy demand in Colombia, from December 15, 2015 to December 31, 2008, provided by the department of energy demand forecasts at the XM subsidiary of ISA, as a public information. Three models were compared to determine the best one: SARIMA model, econometric model with exogenous variables, and Gaussian Bayesian regression model [8], [9], [10].

Figure 1 shows the behavior of the daily electrical energy demand, during a month.

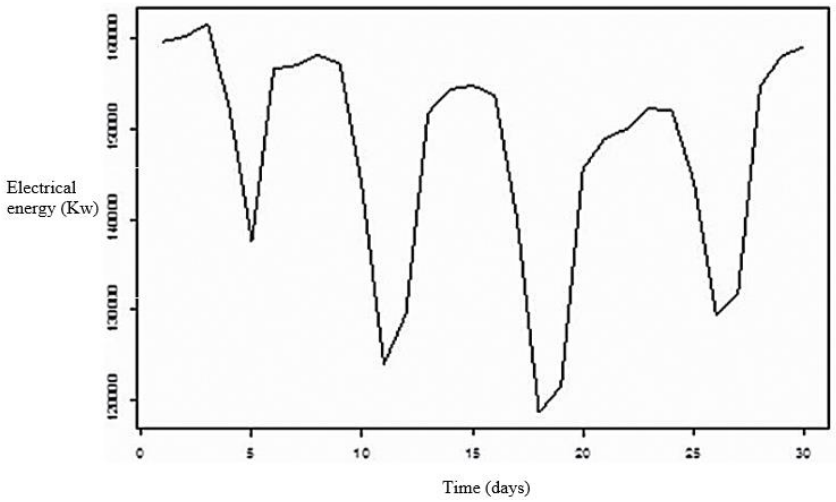

Figure 1. Daily electricity demand in a month

According to figure 1, a very similar behavior can be observed every 7 days, with high demand at the beginning of the week and low demand at the end of the weekend. In this way, a seasonal behavior 7 can be identified. In addition to the above, from interviews with the expert personnel of the company, this hypothesis was tested and other variables that can have important impacts were determined, such as temperature and energy level, information that was useful to estimate the econometric model with exogenous and endogenous variables, as well as the SARIMA model to make the comparison of the three models.

This also served as an input to incorporate some important explanatory variables in the Gaussian Bayesian MCMC regression model.

2.1 Econometric model

The model considered the endogenous variable, Zt , electricity consumption, and explanatory variables such as: endogenous with Zt-k delays, daily temperature, and other exogenous variables such as dummy variables to include interventions in the series (such as: holidays, day of the week), seeking to estimate a model of the form [11], [12]:

$$Zt = \beta 0 + \beta 1 Zt - 1 + \ldots . + \alpha t + \alpha 1 \text{sen}(2\pi t/7) + \alpha 1 \cos(2\pi t/7) + \ldots . + \text{day week} + \text{variables artificials} + \varepsilon t \quad (1)$$

Others were explored in the process, adding or deleting variables if they were significant or not at the 5% level.

2.2 Markov Chain Bayesian Gaussian Regression Model

The general Gaussian Bayesian regression model used was [13], [14]:

$$Zt = \beta 0 + \beta 1 \, Zt - 1 + \ldots . + \alpha t + \ldots . + \text{day week} + \varepsilon t \quad (2)$$

Where:
Zt is the demand for electrical energy
Zt -1 is the variable with delay of the demand of electric energy.
t is the time trend

Day of the week: these are indicator variables (0/1) according to the day, so that they are added to the constant β0 for the respective day.

The statistics used for the estimation of each coefficient corresponds to the mean value of the MCMC sampling performed.

2.3 Final evaluation criteria of the models

In the analysis of the models, hypothesis tests were used to analyze the significance of the parameters of the classical models, based on the assumptions of normality, tests to determine the fulfillment of the assumptions in the residuals. These were not the same for Bayesian processes, where it was necessary to determine whether the simulation generated independent samples, as expected. However, another indicator was considered as the criterion for the final choice of the best model. The ASM indicator was then measured, of two types: adjustment, with data used for estimation and prognosis; with data not used for estimation. Since the final objective was to find a model with adequate capacity for forecasting, it was considered an adequate indicator of success capacity in the models and, therefore, the final criterion of choice [15], [16], [17].

3. Results

The estimates of the statistical models used are presented with the respective characterization of Colombia's daily energy consumption, using the validation tests and comparison criteria necessary for an adequate analysis and selection of the one that allows optimizing in a more integral way, the forecast by showing a lower level of relative absolute error.

3.1 SARIMA Model

When analyzing the autocorrelation values of the series of energy consumption, a seasonality of order s=7 was evident, since in the periods 7 and 14 it took very high values (0,7587 and 0,7147 respectively). Likewise, the autocorrelation of order 1 is very high (0.47852). This indicates that the most obvious delays that have effects on the series are Zt-1, Zt-7 and Zt-14. In addition to these evidences, the automatic R method was used to better detect the parameters indicated for the model. The adequacy of the analysis will be tested with this model. Changes are then introduced using the econometric model. After estimating several models, ARIMA (1,1,1)X(2,0,1) was found with the lowest ASM value. Table 1 shows the SARIMA coefficients.

Table 1. SARIMA Coefficients

Coefficients				
ar1	**ma1**	**sar1**	**sar2**	**sma1**
0,5012	-0,94785	0,97523	0,0085	-0,8078
0,0395	0,017852	0,047852	0,0412	0,02963

Confidence intervals at 95% were estimated using the standard error from the last line of Table 1. The only non-significant coefficient is the second seasonal term. However, by eliminating it, the model loses success capacity, resulting in a better ASM than the one presented in Table 1. The SARIMA model (Table 1) did not adequately meet the assumption of normal residuals according to Jarque Bera, but it did meet the assumption of their uncorrelation, according to the Ljung Box test. In addition, according to the Levene test, the residuals are heteroscedastic. In addition, the adjustment has an ASM of 3.12% and a forecast ASM of 7.30%. Therefore, it was not very wise to use this model to forecast daily energy

consumption. This result indicated that there are important sources of variation, which must be considered before estimating the SARIMA model, aspects considered for the econometric model.

3.2 Econometric model

The following equation reflects the final estimated model, which was improved using 5% significance tests.

$$Zt = \beta 0 + \beta 1\ Zt - 1 + \beta 2\ Zt - 7 + \ldots. + \alpha 1\ t + \alpha 2\ t\ 3 + \ldots. + \beta i\ *\text{Indicators(day week)} + \beta j\ *\text{Indicators levels} + \beta k\ *\text{Temperature} + \varepsilon t \quad (3)$$

In the econometric model estimation, variables such as:

- Time, to determine trend impacts.
- Day, given that the seasonal period in the series is 7.
- Sinusoidal behavior, to study the improvement of seasonality.
- Three lags (past variables), due to the autocorrelation detected in the series.
- Target temperature.

The artificial variables explored were:

- In the variable "Level", demand was catalogued in 4 intervals (1, 2, 3, 4) from lowest to highest respectively. These intervals have a range of plus or minus 14000 MWh each.
- In the variable Level 2, the special dates were catalogued, which are: December 24, 25 and 31. This was done for the model to identify the peaks of Colombian energy demand.
- Level 3 identified the lowest peaks in the data; generally represented it on Sunday.

These levels were incorporated into the model in order to capture peaks and atypical data when forecasting demand. The coefficients of the final model are shown in the second column of Table 2.

Table 2. Final Estimated Coefficients of the final econometric model.

Coefficient	Estimate	Standard error	Value t	Pr(>|t|)
Constant	2,85E+02	9,12E+01	36,147	< 2e-16
t	8,14E-02	1,32E-01	7,298	1,54E-12
t3	-4,77E-06	1,14E-05	-3,75	0,000178
Day Monday	3,95E+00	8,58E-01	4,698	6,12E-06
Day Tuesday	7,47E+00	1,16E+00	6,325	1,88E-10
Day Wednesday	5,36E+00	9,28E-01	5,785	6,25E-08
Day Thursday	4,74E+00	7,89E-01	5,657	2,03E-07
Day Friday	4,96E+00	8,67E-01	4,365	7,78E-07
Day Saturday	1,47E-01	7,55E-01	0,184	0,847106
Temperature	2,36E-01	8,47E-02	3,687	0,000785
Indicator(level)2	1,85E+01	1,47E+00	16,142	< 2e-17
Indicator(level)3	3,69E+01	2,36E+01	18,369	< 2e-17
Indicator(level)4	3,47E+01	1,98E+01	23,786	< 2e-17
Indicator-m(Level3)	1,36E+01	9,78E-02	13,355	< 2e-17
Indicator-s(Level3)	1,96E+01	9,36E-02	20,474	< 2e-17
$\sqrt{Z_{t-7}}$	3,58E-02	2,98E-02	1,966	0,07899
$\sqrt{Z_{t-2}}$	8,69E-02	1,78E-02	5,987	2,21E-07

Hypothesis tests to contrast the significance of quantitative variables were performed by analyzing the P value of the last column of Table 2, of the statistic t, indicating with lower values of the 5% significance level, that the parameter had significance. For the qualitative explanatory variables (artificial or indicator) it was evident that more than one of their coefficients had a P value of less than 5%. This was also checked with the ANOVA type III table, which showed F test p-values of less than 5%.

3.3 Gaussian regression model via Monte Carlo by Markov Chains

A Monte Carlo simulation was carried out with 12896 iterations, and the first 2500 samples obtained by means of the MCMC Regress statistical package were burned in order to guarantee independence in the simulation, which will be tested with the Ljung-Box test. Table 3 shows the mean and standard deviation for each variable, with the standard error. The average was the value used to make forecasts.

Table 3. Estimated Coefficients

	Mean	Standard Deviation	Previous Standard Error (Naive)	Time Series SE
Intercept	-6,55E+05	6,85E+04	6,85E+02	6,98E+04
T	3,76E+02	4,87E+01	3,36E+01	4,36E+01
Day Monday	3,83E+03	8,55E+01	9,65E+01	8,47E+01
Day Tuesday	3,31E+03	6,07E+01	6,47E+01	6,47E+01
Day Wednesday	3,28E+03	6,77E+01	6,65E+01	6,69E+01
Day Thursday	3,19E+03	6,78E+01	6,74E+01	6,65E+01
Day Friday	2,15E+03	8,68E+01	6,36E+01	8,75E+01
Day Saturday	1,33E+03	6,56E+01	8,74E+01	6,36E+01
Zt-1	5,96E-02	3,22E-01	3,22E-03	3,47E-03
sigma2	2,57E+06	1,27E+05	2,56E+03	1,11E+03

The estimation of the model, using MCMC, allowed to simulate the regression parameters, and for each, 10000 data were generated, showing the basic statistics in Table 3. The mean is the estimator chosen as the coefficient to be incorporated in the MCMC regression model, using Gibbs' sampler. It is now necessary to carry out a test to indicate that the simulation results did not show any dependency. For this purpose, the Ljung Box test was applied. Table 4 shows that the samples of the parameters that were simulated are uncorrelated, since the p-values are above 6%. Furthermore, the error indicators obtained are: MAPE adjustment=1.93% and the forecast MAPE=4.95%.

Table 4. Uncorrelation tests of simple parameters

Ljung Box Test	
Variable	Valor P
Constant	0,5587
T	0,2965
Day Monday	0,6847
Day Tuesday	0,1258
Day Wednesday	0,4554
Day Thursday	0,2968
Day Friday	0,6478
Day Saturday	0,06225
Delay Zt-1	0,2578
Sigma2	0,1014

When synthesizing all the models, it was observed that the econometric model with residual SARIMA was the most assertive to forecast this type of demand; this is due to the fact that this technique allowed incorporating exogenous variables, admitting some indicators as level variables that led to a good estimation of atypical data. The Bayesian technique cannot be left aside, which through MCMC simulation made it possible to estimate a model with which forecasts were made with a MAPE value of 4.6%, lower than that of the SARIMA model. Model that can be updated with new data, running the simulation again at each time t.

The Bayesian technique proved to be an alternative, when it is not necessary to comply with premises or even when there are not much historical data.

4. Conclusions

Among all the estimated and analyzed models, it was found that the econometric model with SARIMA errors was optimal to achieve the lowest forecast error, getting closer to the reality of daily energy behavior. This is due to the fact that when incorporating significant variables such as delays, fictitious or level variables, they better explained the structural changes in the series, thus providing a tool that can facilitate the final optimization of Colombia's daily energy supply, due to the confidence of the forecast. The SARIMA model for energy consumption, which was one of those used to estimate the forecast of energy demand by the National Dispatch Center (NDC), showed a problem with the required validation of assumptions and was also surpassed by the efficiency of the econometric model, which suggests rethinking it, with the tool provided in this study.

A great advantage of the Bayesian Gaussian regression method by MCMC is that it does not require the same assumptions as a classical model, and facilitates the updating of new data with the simulation for each time t+1 that is forecast. In this research, the method simulated the parameters of a regression model, generating them without time dependence. In addition, an acceptable prognostic efficiency was shown, providing an alternative when few data and models are difficult to adjust.

References

[1] Castellanos Domíngez, M. I., Quevedo Castro, C. M., Vega Ramírez, A., Grangel González, I., & Moreno Rodríguez, R. (2016). *Sistema basado en ontología para el apoyo a la toma de decisiones en el proceso de gestión ambiental empresarial*. Paper presented at the II International Workshop of Semantic Web, La Habana, Cuba. http://ceur-ws.org/Vol-1797/

[2] Pretnar, A. The Mystery of Test & Score. Ljubljana: University of Ljubljana. Retrieved from: https://orange.biolab.si/blog/2019/1/28/the-mystery-of-test-and-score/ (2019).

[3] Yasser, A. M., Clawson, K., & Bowerman, C.: Saving cultural heritage with digital make-believe: machine learning and digital techniques to the rescue. In Proceedings of the 31st British Computer Society Human Computer Interaction Conference (p. 97). BCS Learning & Development Ltd. (2017).

[4] Khelifi, F. J., J. (2011). K-NN Regression to Improve Statistical Feature Extraction for Texture Retrieval. *IEEE Transactions on Image Processing*, 20, 293-298.

[5] Abdul Masud, M., Zhexue Huang, J., Wei, C., Wang, J., Khan, I., Zhong, M.: Inice: A New Approach for Identifying the Number of Clusters and Initial Cluster Centres. Inf. Sci. (2018). https://doi.org/10.1016/j.ins.2018.07.034

[6] Martins, L.; Carvalho, R.; Victorino, C.; Holanda, M.: Early Prediction of College Attrition Using Data Mining. 16th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1075-1078 (2017)

[7] IHOBE. (1999). *Guía de Indicadores Medioambientales para la Empresa*. Berlin: Ministerio Federal para el Medio Ambiente, la Conservación de la Naturaleza y la Seguridad Nuclear.

[8] Russell, S.; Norvig, P.: Artificial Intelligence A Modern Approach. Pearson Education 3rd Ed, pp. 705 (2010)

[9] Makhabel, B.: Learning Data Mining with R. Packt Publishing 1st Ed, pp. 143 (2015)

[10] Witten, I.; Frank, E.; Hall, M.; Pal, C.: Data Mining Practical Machine Learning Tools and Techniques. Elsevier 4th Ed, pp. 167-169 (2016).
[11] Bishop, C. (1995). Extremely well-written, up-to-date. Requires a good mathematical background, but rewards careful reading, putting neural networks firmly into a statistical context. *Neural Networks for Pattern Recognition*
[12] Gaitán-Angulo, M., Viloria, A., & Abril, J. E. S. (2018, June). Hierarchical Ascending Classification: An Application to Contraband Apprehensions in Colombia (2015–2016). In Data Mining and Big Data: Third International Conference, DMBD 2018, Shanghai, China, June 17–22, 2018, Proceedings (Vol. 10943, p. 168). Springer.
[13] Sanchez L., Vásquez C., Viloria A., Cmeza-estrada (2018) Conglomerates of Latin American Countries and Public Policies for the Sustainable Development of the Electric Power Generation Sector. In: Tan Y., Shi Y., Tang Q. (eds) Data Mining and Big Data. DMBD 2018. Lecture Notes in Computer Science, vol 10943. Springer, Cham.
[14] Perez, R., Inga, E., Aguila, A., Vásquez, C., Lima, L., Viloria, A., & Henry, M. A. (2018, June). Fault diagnosis on electrical distribution systems based on fuzzy logic. In *International Conference on Sensing and Imaging* (pp. 174-185). Springer, Cham.
[15] Perez, Ramón, Carmen Vásquez, and Amelec Viloria. "An intelligent strategy for faults location in distribution networks with distributed generation." *Journal of Intelligent & Fuzzy Systems* Preprint (2019): 1-11.
[16] Bucci, N., Luna, M., Viloria, A., García, J. H., Parody, A., Varela, N., & López, L. A. B. (2018, June). Factor analysis of the psychosocial risk assessment instrument. In International Conference on Data Mining and Big Data (pp. 149-158). Springer, Cham.
[17] Chakraborty, S., Das, S.: Simultaneous variable weighting and determining the number of clusters—A weighted Gaussian algorithm means. Stat. Probab. Lett. 137, 148–156 (2018). https://doi.org/10.1016/j.spl.2018.01.015.