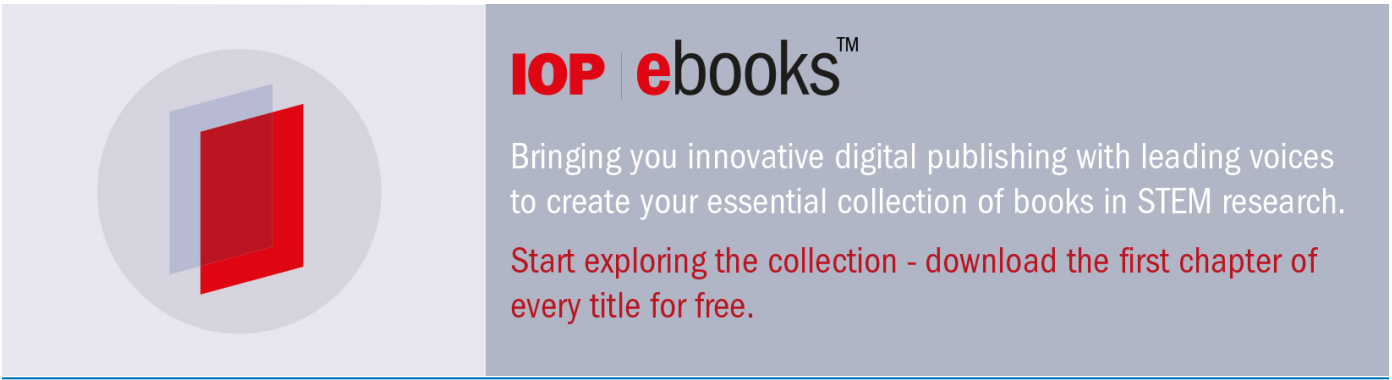
Journal of Physics: Conference Series

PAPER • OPEN ACCESS

Temporary Variables for Predicting Electricity Consumption Through Data Mining

To cite this article: Jesús Silva *et al* 2020 *J. Phys.: Conf. Ser.* **1432** 012033



Temporary Variables for Predicting Electricity Consumption Through Data Mining

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

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

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

Abstract. In the new global and local scenario, the advent of intelligent distribution networks or Smart Grids allows real-time collection of data on the operating status of the electricity grid. Based on this availability of data, it is feasible and convenient to predict consumption in the short term, from a few hours to a week. The hypothesis of the study is that the method used to present time variables to a prediction system of electricity consumption affects the results.

1. Introduction

The short-term prediction is closely related to the problem of consumption peaks, in which a strong increase in demand is observed in a short period of time. This event results in a serious problem that, if not properly predicted and managed, causes economic losses, damage to equipment at different levels, and service outages or price increases (both as a way to discourage consumption in economies that allow consumption by time slots, and due to the need to repair or replace damaged equipment) [1], [2], [3]. Thus, this prediction plays an important role for companies in the electricity sector, especially in terms of distribution, because it allows a better outlook for making strategic and operational decisions [4].

The demand of energy is then a subject of great importance but of difficult solution for different reasons. Since electrical demand is the sum of all individual consumptions of the nodes in the network, determining the levels of electrical demand is a non-stationary and random process, composed of many and diverse individual components. The factors that affect it can be classified into economic, temporal, climatic and random effects, both separate and compound [5]. The behavior of the system is also strongly affected by the scale at which the measurement is made, being softer as the level of aggregation of consumers increases. Electricity consumption at city level or above is reasonably stable, while the electricity consumption of a small group of consumers or isolated houses is a strongly random and noisy process [6], [7].

There are several studies that address the problem of determining the necessary input variables for STLF. For selecting the most significant variables, [8] classifies a group into groups with "soft" limits based on similarities with others, without abrupt transitions between groupings. This property allows an element to belong to more than one group. These sets are used as inputs for the prediction system. On the other hand, for selecting the input variables, [3] use the maximum conditional entropy, which determines the relevance of each variable in the prediction of electricity demand, selecting the best ones. However, in [9], the phase-space embedding method is used to identify a set of "independent" variables

that are related to the time series of electricity demand. In the mentioned cases, the methods were used to determine the input variables in systems based on Artificial Neural Networks.

In [10], a hybrid PSO-SVR system is proposed, using a Particle Swarm Optimization (PSO) algorithm for selecting the optimal input variables, limiting the number of input variables in Support Vector Regression (SVR) for STLF. On the other hand, [2] make the selection of variables. At the same time, [11] states that in the selection of variables the best characteristics do not necessarily lead to a good selection, and that part of the information provided by the eliminated variables is lost. Therefore, the authors use the Random Forest Method to reduce the set of input variables.

Focusing on the temporal issue, there are several methods to represent date-related variables used in electricity consumption prediction models. Among the most commonly used are several variations of DOY (Day-Of-Year), such as the sample number, sequential numbering and non-linear variables derived from the date. Based on this, the study introduces the hypothesis that the method used to include time variables in a prediction system of electricity consumption influences the results.

A comparison of different methods of representation of temporal variables is then proposed to determine which is the most convenient for the prediction of short-term electricity consumption. As a case study, the electricity consumption of the city of Medellín in Colombia, period 2016 to 2018, was chosen. A simple linear regression method is used to validate the hypothesis, since the objective of this research is to determine the impact on the way in which time variables are presented and not to construct a prediction system.

2. Method

The purpose of the research is to propose as a hypothesis that the inclusion of time variables affects the quality of the prediction of electricity consumption. To validate it, a comparison between the results of the STLF experiments is performed using different methods to represent temporal variables. The analysis is based on real data from the city of Medellín in Colombia.

Based on the recurrent use of the *DOY* variable, this approach is used as the first method and starting point for establishing the first point of comparison. As a second method, a separation of the date variable was proposed, breaking it down into four individual variables: year, month, day, and day of the week. Finally, an adaptation of the method described in [12] is used, where the author uses a pre-processing (non-linear operations) on the input time variables. The latter approach is similar to the previous one, as it denotes an addition of time variables derived from the originals.

Thus, three groups of variables were formed for the different tests detailed in Table 1. The fourth group corresponds to the climatic variables and variables with information on the atypical days that will be used in all the tests. The combination of these three methods is then tested following what was done by [13].

Since electricity consumption continuously varies over time, it is feasible to apply different techniques and time series methodologies to predict consumption based on available historical data. One of these techniques consists of using current and/or past information to predict future variables [14]. This case study proposes the use of data from previous days, related to both the demand for electricity and climate variables of the region and data related to dates.

At the same time, in order to improve the prediction precision, there are variables whose future values can be known in advance. This is the case of date type variables and related data (holidays, public holidays or regional special events). It should be noted that the presence of atypical days such as weekends and holidays influence electricity consumption [15]. For this reason, this information was added as common input variables for all tests. The advantage of the possibility of adding information of the prediction day is that it produces a decrease of the error that can be generated by possible changes of tendencies that are reflected or altered by the presence of this type of information.

With regard to data separation, since it is a time series, the validation set consisted of 30% of the last samples in the series consecutively, i.e. the most recent samples. In turn, the remaining 70% was used for training. As part of the tests that were carried out, the variable separation procedure described in [16] was used, which is closer to the actual use of electricity companies. However, the improvements found are not significant.

In order to evaluate the results obtained with the selected model, the error obtained through different metrics is analyzed: Root Mean Squared Error (RMSE), that is expressed in the same unit of measurement than the variable to be estimated, facilitates its interpretation; Relative Error (RE) expressed as a percentage; Mean Bias Error (MBE) that allows the analysis of whether there is an underestimation or an overestimation in the prediction of electricity consumption; and Pearson's Linear Correlation Coefficient (R) that helps determine the degree to which the data follow the general trend of the model.

Table 1. description of the used database.

Group	Origin	Minimum	Maximum	Average	STD
1	Date	1	366	188,5	107,38
2	Date	1 1	12 31	6,70 15,72	3,52 8,82
3	Date	0 0	1 1	0,04 0,04	0,19 0,19
	Atypical days	0,00	1,00	0,28	0,45
		0,00	1,00	0,04	0,19
	Climate - region 1	0,01	600,25	214,56	151,52
	Climate - region 2	0,01	655,36	227,67	145,8
4		12,25 -6,7 0 0	2007,04 25,6 655,36 989,4	792,76 12,86 209,27 964,32	347,88 6,62 153,82 29,57
	Climate - region 3	0	635,04	208,94	153,8
	Climate - region 4	0	767,29	243,97	155,3
5	Climate - region 5	3615,78	10644,06	5907,13	875,46

2.1 Data used

For the development of the tests, information of databases from different sources was used. Climate data were collected from 5 different meteorological stations, belonging to the city's Agroindustry Experimental Station. Electricity consumption data, on the other hand, were provided by ¨Empresa de Distribución Eléctrica¨ of Medellín and correspond to measurements of the demand for electricity consumption in each province. Finally, data on dates and holidays in the Republic of Colombia were acquired through the website of the National Directorate of Political Affairs, which reports to the Ministry of the Interior and Transport.

The data used correspond to periods from 01-01-2016 to 29-12-2018 and contains daily samples of the variables detailed in it. In addition, the minimum and maximum temperatures are added to the square belonging to each of the stations in a similar way to that observed in [17].

Preliminary tests determined that the cubic variables do not have a major influence on this case. In the initial data analysis procedure, and as is usual in distributed sensor networks, records were found with zero or missing values in both climate and electricity consumption data. This is possibly due to problems in measuring equipment or information acquisition processes. Due to the fact that the amount of data affected is not significant, it was decided to carry out a pre-processing process through which records that present an anomaly were eliminated. The percentage of cells with missing data is 9.5%, resulting in 430 usable samples out of a total of 4521.

2.2 Linear Regression

Linear Regression is one of the most widely used statistical methods, partly due to its easy and simple interpretation of the model to be acquired. Due to such characteristics and considering that the objective of the present study is to validate the hypothesis, not to obtain the best possible prediction, this technique was used to estimate the values of electrical consumption, in the same way as [13], [18], [19],

understanding that for definitive versions of the system it is necessary to use Weighted or Multivariate Linear Regression [20].

It is important to make clear the fact that in the problem addressed, many of the input variables have a linear correlation, so that most of the linear systems involved in the work are poorly conditioned. This type of linear systems produces a strong variation in the output to small changes in the input, making the solution not adequate in many cases. To solve this problem, Moore-Penrose pseudoinverse [21] is used, since it is capable of offering good solutions in the case of badly conditioned systems.

3. Results

In the tests that were carried out, daily samples were available, with the objective of predicting one day forward (One-Day Ahead) on the variable electricity consumption. To improve the estimate, in addition to taking current samples, samples from several consecutive previous days or delays (auto-regressive time series) are considered.

In Figure 1, the error variation is shown using the DOY and atypical days variables, and the climatic variables, for the 4 types of errors used as comparison metrics. It can be observed that the best result is achieved with a Delay 2, i.e. considering current samples and those of 1 day backwards. However, for higher delay values, the training error decreases while the validation error tends to increase. This is due to the fact that the increase in variables that implies a large delay causes an increase in the complexity of the linear regression model, which in turn generates an overfitting of the system, as described in [19].

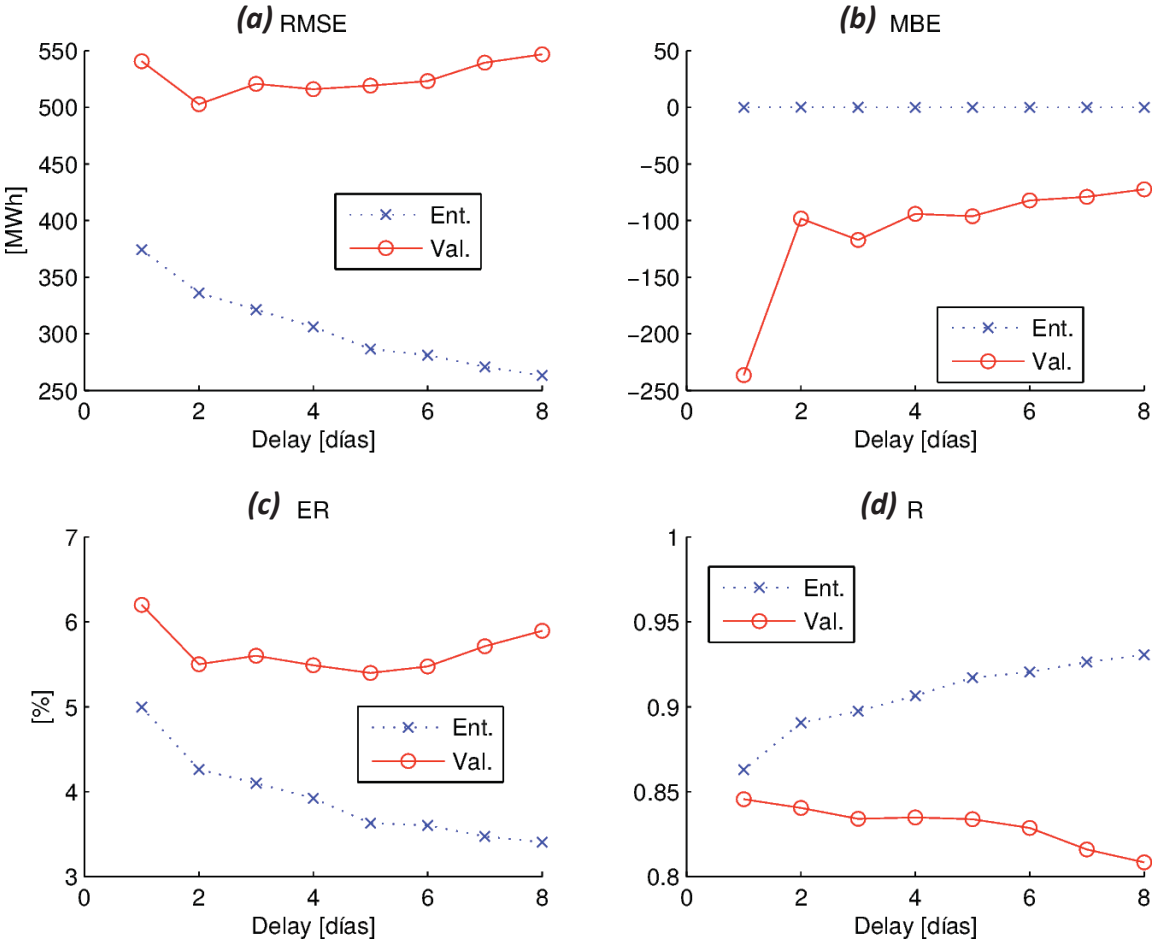


Figure 1. Errors obtained for different delay values

Table 2 and 3 detail the errors obtained at provincial level, for the training and validation sets respectively. It can be noted, both in training and in validation, that the best result was obtained by using the variables of groups 2-4 as system input. A 12.45% improvement was achieved in the Relative Error with respect to the non-use of time variables. The other alternatives do not bring a significant improvement.

Table 2. and Validation Errors

Groups of variables	**Description**	**SSE**	**MBE**	**R**	**RMSE**	**Err.Rel**
4	Basic Variables	189564872,4	0,0871	0,8999	357,1528	4,34
2-4	Decomposition into dates	136542927,4	0,0000	0,9125	311,8654	3,98
3-4	Non-linear components	178524556,8	0,0125	0,8587	351,2478	4,35
1-4	DOY	136587421,6	0,0311	0,8698	334,0144	4,69
1-2-3-4	All	1289657445,6	0,0001	0,9147	310,5541	3,58

Table 3. Training and Validation Errors

Groups of variables	**Description**	**SSE**	**MBE**	**R**	**RMSE**	**Err.Rel**
4	Basic Variables	142451612,6	-84,9913	0,8789	511,5247	5,49
2-4	Decomposition into dates	187544256,0	4,0789	0,8787	464,4571	4,92
3-4	Non-linear components	145784450,8	-131,325	0,8958	513,872	5,89
1-4	DOY	141985477,4	-99,3245	0,8783	572,7734	5,47
1-2-3-4	All	119863205,7	11,0787	0,8789	498,7503	4,58

4. Conclusions

This study proposed the hypothesis that the way in which temporal variables are presented to a prediction system of electricity consumption influences the quality of the results. This hypothesis was validated using different methods to represent these variables, applying it together with climate data and electricity consumption of the province of Tucumán. The division of the temporal variable in day, day of the week, month and year in individual form for each period involved in the problem, turned out to be the most convenient method, obtaining an improvement of up to 12.45% with respect to other methods considered.

The use of all available time variables together led to a reduction in training and validation errors. This implies that none of the time variables introduce noise into the prediction for the data set used. It is desirable to generate simpler models that depend on fewer variables and have the same level of error. This is because excessively complex models have a good ability to adjust training data, but may not adapt well to new data (overfitting) [19]. Therefore, applying variable selection techniques will help reduce model complexity by eliminating unnecessary, redundant and noisy variables.

It should be noted that the study focused on comparing the use of different sets of variables without giving too much importance to the prediction method. For this reason and for simplicity, linear regression was used, obtaining an error of 4%, which means that the problem is linear.

It is important to note that the scale at which the analysis is conducted can also influence the results, causing the ideal way in which the data is presented to change.

References

[1] R. Sevlian and R. Rajagopal, "Short Term Electricity Load Forecasting on Varying Levels of Aggregation," ArXivPrepr.ArXiv14040058, 2014.
[2] Z. Y. Wang, C. X. Guo, and Y. J. Cao, "A new method for short-term load forecasting integrating fuzzy-rough sets with artificial neural network," in Power Engineering Conference, 2005. IPEC 2005. The 7th International, 2005, pp. 1–173.
[3] I. Drezga and S. Rahman, "Input variable selection for ANN-based short-term load forecasting," IEEE Trans. Power Syst., vol. 13, no. 4, pp. 1238–1244, Nov. 1998.
[4] Y.-C. Guo, "An integrated PSO for parameter determination and feature selection of SVR and its application in STLF," in 2009 International Conference on Machine Learning and Cybernetics, 2009, vol. 1, pp. 359–364.
[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] 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.
[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] 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
[17] 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*.
[18] 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).
[19] 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/

[20] 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).

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