Prediction of monthly electricity consumption by cantons in Ecuador through neural networks: A Case Study

Jorge I. Guachimboza-Davalos[1], Edilberto A. LLanes-Cedeño[1], Rodolfo J. Rubio-Aguiar[1], Diana B. Peralta-Zurita[1], and Oscar F. Núñez-Barrionuevo[1]

Universidad Internacional SEK, Quito-Ecuador
jiguachimboza.mee@uisek.edu.ec

Abstract. The present work shows a multivariate regression through deep learning related to energy consumption in the province of Pichincha-Ecuador in relation to its cantons. This is done to establish consumption habits in relation to the existing population. For this, a data preprocessing stage was implemented and later an artificial neural network of a hidden layer and 15 neurons. As a result, a mathematical model with an absolute error of 2.2 % is obtained.

Keywords: power consumption · multivariate regression · neural network

1 Introduction

Currently, one of the biggest government concerns is global warming. Since this causes an abrupt change in the environmental conditions of the planet, altering the living conditions of people, animals and plants [1]. This problem is mainly caused by the excessive emission of polluting gases (carbon dioxide CO_2, carbon dioxide CO, hydrocarbons, oxides of nitrogen, sulfur, among others) of companies that use fossil fuels for the development of a product or service [2,3]. Related to this, is the production of emissions due to the electricity demand of the population. The same, which is a key element for the socio-economic development of each country. That said, 41% of the world's CO_2 emissions are attributed to electricity generation. The same ones that contribute the majority (around 99.5 %) of CO_2.

In Latin America, there is an average of 0.65 tons of CO_2 for each MWh generated. [4] [5]. In Ecuador, this result is 397.5g of CO_2 per KWh, where the highest value corresponds to 2010 since the generation of energy from non-renewable sources was (52.2 %) [6]. Currently, with the change in the production matrix, 60.85 % of electricity production is based on renewable energy. Where 58.53 % is related to Hydroelectric sources [7]. This production is oriented 30.93 % to residential consumption and 26.01 % to the Industrial sector [6]. With this, these forms of use have an important factor for the projection of electrical consumption and the planning of new sources of electricity generation [8]. However, the use of this resource in homes and industries show a high variation that depends on the lifestyle of the users and the type of good or service that is generated [9]. For this reason, it is necessary to implement prediction models that allow searching for people's consumption patterns in relation to a variety of attributes (consumer characteristics). However, these studies have been focused on consumption by independent households with the growth of the Internet of Things. Therefore, there are no adequate metrics for consumption trends organized by types of home users (rural and rural) or from different provinces of a country or region. In addition, The main problem to obtain a reliable forecast of Electric power consumption is choosing a forecast model that accurately presents initial data. The demand pattern is very complex due to the deregulation of the energy markets. Therefore, finding an appropriate forecasting model for a specific power grid is not an easy task. Although many forecasting methods were developed, none can be generalized for all demand patterns.

To do this, machine learning algorithms allow mathematical-computational calculations to be made to extract knowledge from the data. This criterion is widely used since it seeks to emulate certain functioning of the human brain for decision-making based on experience. However, data analysis has its limitations due to the poor structure of information storage by government entities. This is generally caused by the use of categorical variables (they do not contain numbers). That is why it is necessary to code these variables to numerical form. However, this results in increasing the size of the database as it uses ones and zeros to convey the necessary information and can confuse the learning model by providing non-existent information to the model. However, with an adequate methodology, sub-processes of data cleaning, normalization, selection of characteristics, among others, can be carried out. As a result, the mathematical model can learn from true information and dismiss the variables that are of lesser contribution [10]. With this, consumption parameters can be obtained and the future demand for electricity needed to satisfy the increase in this service can be forecast in the future.

However, there are classic approaches to machine learning that can represent the study of a phenomenon but that depends on the type of data to be used (categorical and numerical). This learning model requires feature extraction and a monitoring algorithm to generate a model. On the other hand, deep learning is based on the use of neural networks, since it allows the aforementioned stages to be carried out in a computational model. This new approach is based on neural networks [11]. This criterion allows you to organize what you have learned (self-organization). Therefore, they have the ability to be fault-tolerant, since the more times it comes into operation (iterate) it can solve its error (update of weights). For this reason, you cannot experience sudden system crashes [12].

Some works such as [13,14] present systems for measuring electricity consumption in homes to store and analyze them in the cloud. On the other hand, they do not present a sector-specific solution and are based on personal consumer trends. In this case, the application is focused on the end-user and not on a state-level analysis. It is for this reason that the present work carries out an analysis of the different consumption by cantons of the Pichincha province in Ecuador with the objective of determining parameters and consumption trends to present proposals for energy demand towards the future. For this, the information from the Electricity Regulation and Control agency has been compiled to implement deep learning through the use of neural networks. For this, a preprocessing is carried out with processes for the selection of characteristics and normalization of the database. Subsequently, an analysis of the activation functions in the different layers of a neural network is performed to determine the appropriate model that can provide the best prediction of desired values of electrical consumption. For this, different metrics are performed to calculate the forecast error. As relevant results, the model had a forecast error of 2.2 %.

The rest of the document is structured as follows: Section 2 shows the methodological scheme for data analysis. Section 3 presents the analysis of the neural network to provide the best forecast for electricity consumption. Section 4 shows the results obtained with the different error analyses. Finally, section 5 presents the results and future work.

2 Materials and Methods

This section shows, on the one hand, the acquisition of data (section 2.1), preprocessing of the information using the selection of characteristics (section 2.2) and the normalization of the database (section 2.3). In addition, it presents the data analysis scheme (section 2.4).

2.1 Data Acquisition

Data collection is through the information presented by the Ministry of Electricity and Renewable Energy in its monthly accounts during 2019. In these reports, the company that distributes electricity to the provinces, cantons, and parishes is explained in detail. In this specific case, the consumption information for the years 2019 and 2018 is taken. All data is stored in the matrix $\mathbf{Y} \in \mathbb{R}^{m \times n}$ where m represents the number of samples and n the attributes of the database. In this case, $m = 2{,}550$ and $n = 11$. Attributes refer to: Year of consumption, month, a company that bills, province, canton, parish, type of equipment with a voltage of 220, number of customers, energy billed, consumption increase, residential consumption, subsidy energy generated, billing of the service and value of the subsidy.

2.2 Feature Selection

The choice of a training set for the prediction algorithms is a crucial aspect since the response time must be adequate [15]. Typically, for a supervised data analysis task, a large number of attributes can be candidates for characterization purposes. However, many of these attributes may be irrelevant or redundant and consequently, the algorithm may be over-fitting and some important features may be affected. Therefore, the performance or acceptance of the model may not meet what was expected [16]. In fact, in many applications involving large data sets, predictions do not work properly until the unwanted features are necessarily removed [15].

In this sense, the feature selection task reduces the number of attributes that represent a data set that best fits a model or machine learning task. There is a wide range of criteria and algorithms that can be used for this task. However, depending on the nature of the data and the objective of the data analysis, some feature selection criteria may be more appropriate. The most important criteria are taken into account in this work. On the one hand, the correlation of variables is taken into account in a statistical model under the criteria of p_{value} and R^2 with the use of the `backward elimination` criterion. This combination indicates that changes in the predictors are related to changes in the response variable and that the model largely explains the variability of the response. With this, an estimate can be obtained of how the model represents the phenomenon studied [16].

2.3 Database normalization

Index normalization means adjusting the measured values on different scales with respect to one in common. This process is done prior to the process of making mathematical models. This avoids the existence of variables that, according to their nature of the data, contribute a large amount to the model as they have very high values in relation to the rest. In addition, it prevents the use of categorical variables from being diminished due to its low scale (0 to 1). That is why there are different methods and forms of standardization.

2.4 Data analysis scheme

The present work is aimed at finding the best model that fits the study presented where the power consumption of the Ecuadorian population can be prognostic with the most representative variables. The proposed scheme is shown in Fig. 1.

2.5 Neural Network training

The information obtained by the neural network cannot transmit everything learned, it needs an activation function representing an activation potential (a binary form of event probability). In this sense, the activation function allows the

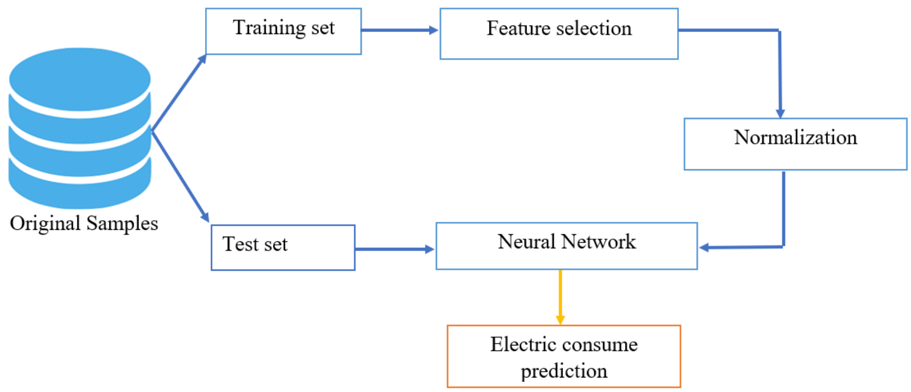

Fig. 1. Data analysis scheme

update of weights (learning of the neuron). The equation 1 shows the operation of a neural network for prediction. Furthermore, it is shown in Fig. 2.

$$Y = f(\sum_{i=0}^{i=m}\sum_{j=0}^{j=n} x_i w_{ij}) \tag{1}$$

3 Results

The results of the present work are divided into: (3.1) selection of characteristics and (3.2) result of the neural network.

3.1 Feature Selection

The codification of the categorical variables in relation to the canton, since by parish the model would have a very high codification and would provide more information (33 rural and 32 urban parishes) than the numerical variables. Consequently, the coding of cations is shown in table 1

In this sense, we proceed to eliminate these variables in order to improve the model. Fig. 4 shows the result of the analysis of covariance where it indicates that the model fit has not changed (R squared) when eliminating the variables previously mentioned.

3.2 Neural Network

Once the variables have been selected to implement them in the neural network. A minimum of 10 neurons is established to train the model and test its

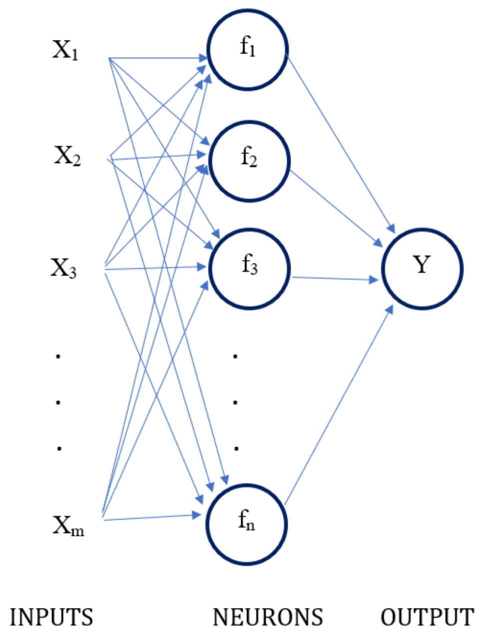

Fig. 2. Neural Network designed

Table 1. Pichincha cantons categorical codification

Cantons	cod1	cod2	cod3	cod4	cod5	cod6	cod7
Cayambe	1	0	0	0	0	0	0
Quito	0	1	0	0	0	0	0
Pedro Vicente Maldonado	0	0	1	0	0	0	0
Puerto Quito	0	0	0	1	0	0	0
Rumiñahui	0	0	0	0	0	1	0
San Miguel de los Bancos	0	0	0	0	0	0	1

operation. As a result, 15 neurons are established to represent a lower forecast dispersion error (MSE) and the absolute mean deviation error (MAD). This value is monetary. Table 2 shows this analysis.

Table 2. Error analysis for numbers of neurons

Neurons	MSE	MAD
10	4.5%	125
12	3.8%	75
15	2.2%	45
17	2.2%	45
20	2.2%	45

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                4.968e+01  2.446e+01   2.031   0.0426 *
ï..Canton2                -2.136e+01  4.352e+01  -0.491   0.6238
ï..Canton3                 8.331e+00  3.431e+01   0.243   0.8082
ï..Canton4                -6.127e+01  4.065e+01  -1.507   0.1322
ï..Canton5                 5.440e+01  6.688e+01   0.813   0.4162
ï..Canton6                 1.893e+02  8.257e+01   2.293   0.0221 *
ï..Canton7                -1.960e+01  9.091e+01  -0.216   0.8294
Equipamiento              -1.672e-01  3.208e-01  -0.521   0.6023
Numero.Clientes           -1.391e+00  8.446e-02 -16.471   <2e-16 ***
Energia.Facturada.k.Wh     1.071e-01  1.294e-03  82.750   <2e-16 ***
Consumo.Incremental.kWh.  -1.232e-02  8.861e-04 -13.902   <2e-16 ***
Consumo.Residencial.kWh.  -7.286e-03  8.133e-04  -8.958   <2e-16 ***
Energia.Subsidio.PEC.kWh.  1.575e+02  1.341e+02   1.175   0.2404
ValorSubsidio.PEC.USD.    -1.750e+03  1.490e+03  -1.175   0.2404
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 269.3 on 818 degrees of freedom
Multiple R-squared:  0.9997,    Adjusted R-squared:  0.9997
F-statistic: 2.186e+05 on 13 and 818 DF,  p-value: < 2.2e-16
```

Fig. 3. Covariance analysis with attributes of matrix **Y**

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               3.251e+01  1.372e+01   2.369   0.0181 *
ï..Canton2               -1.582e+01  4.325e+01  -0.366   0.7147
ï..Canton3                1.156e+01  3.415e+01   0.339   0.7350
ï..Canton4               -5.909e+01  4.064e+01  -1.454   0.1463
ï..Canton5                5.637e+01  6.677e+01   0.844   0.3987
ï..Canton6                1.905e+02  8.253e+01   2.308   0.0213 *
ï..Canton7               -1.220e+01  9.088e+01  -0.134   0.8932
Numero.Clientes          -1.344e+00  7.744e-02 -17.362   <2e-16 ***
Energia.Facturada.k.Wh    1.084e-01  1.048e-03 103.433   <2e-16 ***
Consumo.Incremental.kWh. -1.197e-02  8.430e-04 -14.196   <2e-16 ***
Consumo.Residencial.kWh. -7.690e-03  7.775e-04  -9.890   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 269.6 on 821 degrees of freedom
Multiple R-squared:  0.9997,    Adjusted R-squared:  0.9997
F-statistic: 2.837e+05 on 10 and 821 DF,  p-value: < 2.2e-16
```

Fig. 4. Covariance analysis with attributes of matrix **X**

With this, a prediction of electricity consumption can be established in relation to the payment of the return. Fig. 5 shows the prognosis generated by the neural network.

Finally, a summary of all the parishes is made to have a final summary of the consumption by canton in relation to the year 2019. As can be seen in Fig. 6, the forecast is very accurate for true consumption.

4 Conclusions and Future Work

The present work was able to verify the functioning of learning from neural networks. With this, the prediction of the target variable was very close to the real value. In this sense, it can be seen that by using these criteria, strategies for the future can be defined with the aim of improving electricity service and

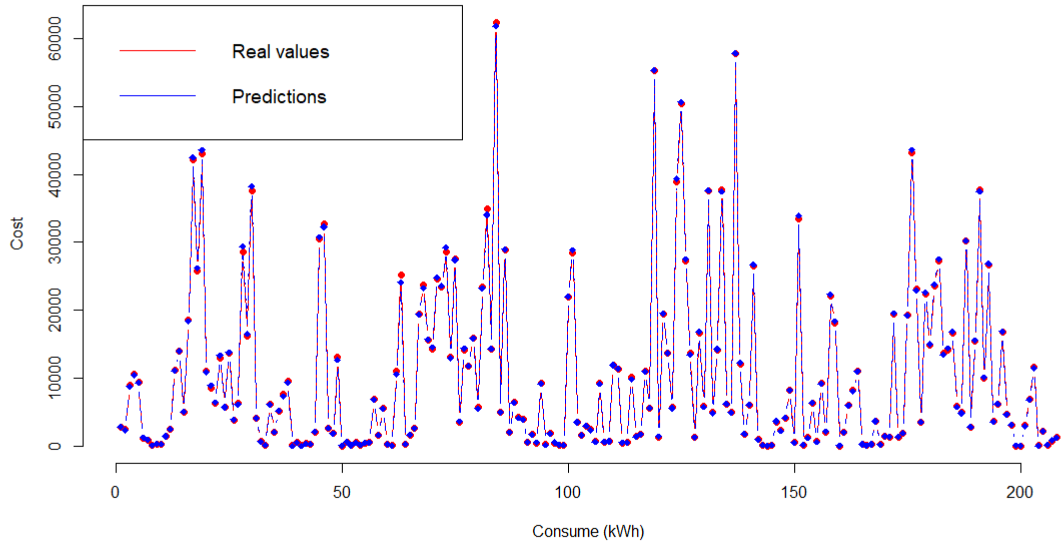

Fig. 5. Cost predictions since neural network analysis

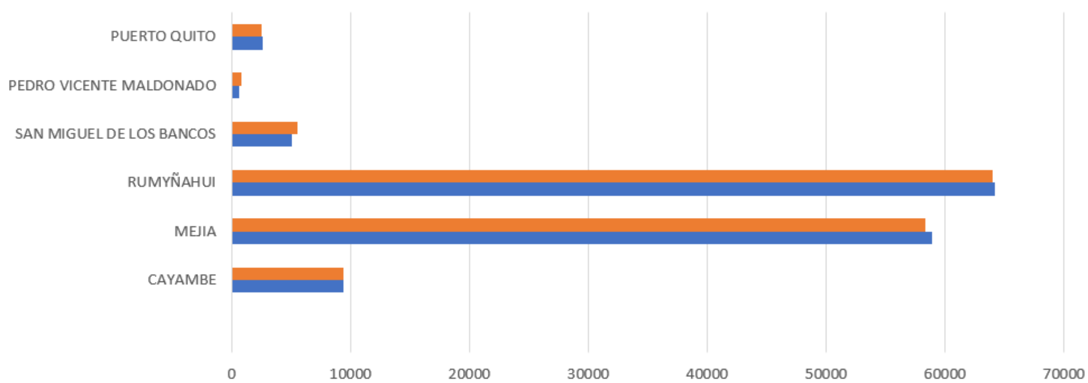

Fig. 6. Final regression model

establishing clear policies for better use of the country's energy resource.

As future works, it is proposed to carry out a deeper analysis of the information in relation to the energy consumption of the country related by parishes.

References

1. X. Ye, X. Wu, and Y. Guo, "Real-time Quality Prediction of Casting Billet Based on Random Forest Algorithm," in *Proceedings of the 2018 IEEE International Conference on Progress in Informatics and Computing, PIC 2018*. Institute of Electrical and Electronics Engineers Inc., may 2019, pp. 140–143.
2. Y. Liu, W. Wang, and N. Ghadimi, "Electricity load forecasting by an improved forecast engine for building level consumers," *Energy*, vol. 139, pp. 18–30, 2017.
3. "UEM en Funcionamiento – Ministerio de Educación." [Online]. Available: https://educacion.gob.ec/uem-en-funcionamiento/
4. X. M. Zhang, K. Grolinger, M. A. Capretz, and L. Seewald, "Forecasting Residential Energy Consumption: Single Household Perspective," in *Proceedings - 17th IEEE International Conference on Machine Learning and Applications, ICMLA 2018*. Institute of Electrical and Electronics Engineers Inc., jan 2019, pp. 110–117.
5. L. Diao, Y. Sun, Z. Chen, and J. Chen, "Modeling energy consumption in residential buildings: A bottom-up analysis based on occupant behavior pattern clustering and stochastic simulation," *Energy and Buildings*, vol. 147, pp. 47–66, jul 2017.
6. "Ministerio de Electricidad y Energía Renovable – Ente rector del Sector Eléctrico Ecuatoriano." [Online]. Available: http://historico.energia.gob.ec/
7. R. Parra Narváez, "Factor de emisión de CO₂ debido a la generación de electricidad en el Ecuador durante el periodo 2001-2014," *Avances en Ciencias e Ingeniería*, vol. 7, no. 2, dec 2015.
8. M. Baquero and F. Quesada, "Eficiencia energética en el sector residencial de la Ciudad de Cuenca, Ecuador," *MASKANA*, vol. 7, no. 2, pp. 147–165, dec 2016.
9. H. X. Zhao and F. Magoulès, "A review on the prediction of building energy consumption," pp. 3586–3592, aug 2012.
10. A. Toapanta-Lema, W. Gallegos, J. Rubio-Aguilar, E. Llanes-Cedeño, J. Carrascal-García, L. García-López, and P. D. Rosero-Montalvo, "Regression models comparison for efficiency in electricity consumption in ecuadorian schools: A case of study," in *Applied Technologies*, M. Botto-Tobar, M. Zambrano Vizuete, P. Torres-Carrión, S. Montes León, G. Pizarro Vásquez, and B. Durakovic, Eds. Cham: Springer International Publishing, 2020, pp. 363–371.
11. X. Wang, C. Lai, D. Yu, Y. Xu, and Y. He, "Application of energy performance contracting in distribution transformer based on neural network prediction," in *2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, March 2019, pp. 42–46.
12. I. Petkovic and N. Balaban, "Detecting defaulters for consumed electric energy with neural network clustering," in *2008 6th International Symposium on Intelligent Systems and Informatics*, Sep. 2008, pp. 1–4.
13. sohrab Rezaei, A. Sharghi, and G. Motalebi, "A framework for analysis affecting behavioral factors of residential buildings' occupant in energy consumption," *journal of Sustainable Architecture and Urban Design*, vol. 5, no. 2, pp. 39–58, 2018. [Online]. Available: http://jsaud.sru.ac.ir/article{_}895.html{%}0Ahttp://jsaud.sru.ac.ir/pdf{_}895{_}9f6e3441f7399851f2185e37696ed98e.html
14. B. J. Johnson, M. R. Starke, O. A. Abdelaziz, R. K. Jackson, and L. M. Tolbert, "A method for modeling household occupant behavior to simulate residential energy consumption," in *2014 IEEE PES Innovative Smart Grid Technologies Conference, ISGT 2014*, 2014.

15. Y. Zhai, W. Song, X. Liu, L. Liu, and X. Zhao, “A chi-square statistics based feature selection method in text classification,” in *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*, Nov 2018, pp. 160–163.
16. M. Peker, A. Arslan, B. Şen, F. V. Çelebi, and A. But, “A novel hybrid method for determining the depth of anesthesia level: Combining relieff feature selection and random forest algorithm (relieff+rf),” in *2015 International Symposium on Innovations in Intelligent SysTems and Applications (INISTA)*, Sep 2015, pp. 1–8.