Multi-criteria methodology based on data science for the selection of the optimal forecast model for residential electricity consumption

Metodología multicriterio basada en ciencia de datos para la selección del modelo óptimo de pronóstico del consumo de energía eléctrica residencial

C. A. Yajure-Ramírez

DOI: https://doi.org/10.22517/23447214.25335
Scientific and technological research article

***Abstract*— There is a wide variety of techniques and models for forecasting electrical energy consumption, depending on both the type of user, the forecast horizon, and the resolution of the available data. Likewise, there are different metrics to evaluate the performance of these models. So, in this research an integrated multi-criteria methodology is proposed to select the best forecast model for residential electricity consumption, using the Analytical Hierarchical Process (AHP) to establish the weights of relative importance of the decision criteria, and the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) to make the selection of the optimal model. The methodology is in turn framed within a data science process, through which the data is extracted, processed, and analyzed, prior to the application of the machine learning algorithms to obtain the forecast models, which will correspond to decision alternatives. The performance metrics in the evaluation phase of the models, and the performance metrics obtained from the forecast phase, are considered as the decision criteria. From the pairwise comparisons technique, it was obtained that the mean absolute percentage error (MAPE) of the prognosis phase was the criterion with the greatest weight of importance, followed by the coefficient of determination R^2 and the MAPE of the evaluation phase. From the TOPSIS method, the Multiple Linear Regression model was selected as the optimal forecast model.**

***Index Terms*— AHP, data science, machine learning, pairwise comparisons, regression, TOPSIS.**

***Resumen*— Existe una gran variedad de técnicas y modelos para el pronóstico del consumo de energía eléctrica, dependiendo tanto del tipo de usuario, como del horizonte de pronóstico y de la resolución de los datos disponibles. Asimismo, existen distintas métricas para evaluar el desempeño de estos modelos. Entonces, en esta investigación se propone una metodología integrada multicriterio para seleccionar el mejor modelo de pronóstico del consumo de energía eléctrica residencial, utilizando el proceso jerárquico analítico (AHP) para establecer los pesos de importancia relativa de los criterios de decisión, y la técnica para el orden de preferencia por similitud con la solución ideal (TOPSIS) para hacer la selección del modelo óptimo. La metodología se enmarca a su vez dentro de un proceso de ciencia de datos, a través del cual se extraen, procesan y analizan los datos, previo a la aplicación de los algoritmos de aprendizaje automático para obtener los modelos de pronósticos, que se corresponderán con las alternativas de decisión. Las métricas de desempeño en la fase de evaluación de los modelos, y las métricas de desempeño obtenidas de la fase de pronóstico, son consideradas como los criterios de decisión. De la técnica de comparaciones pareadas se obtuvo que el error porcentual absoluto medio (MAPE) de la fase de pronóstico fue el criterio con mayor peso de importancia, seguido del coeficiente de determinación R^2 y del MAPE de la fase de evaluación. A partir del método TOPSIS, se seleccionó el modelo de Regresión Lineal Múltiple como el modelo óptimo de pronóstico.**

***Palabras claves*— AHP, aprendizaje automático, ciencia de datos, comparaciones pareadas, regresión, TOPSIS.**

This manuscript was sent on January 11, 2023 and accepted on June 21, 2023.
C.A. Yajure-Ramírez is a postgraduate professor of the master's program in Operations Research at the Basic School of Engineering of the Central University of Venezuela, Caracas, Venezuela (e-mail: cyajure@gmail.com).

I. INTRODUCTION

AN important element for the operation and planning of electrical systems is the forecast of the consumption of electrical energy produced by that system. Within the electricity consumption forecast, some factors must be considered, one of which is the type of customer. That is, whether the consumption data is associated with residential, commercial, official, industrial, or other customers. Additionally, the forecast horizon must be defined. For example, in [1] they classify the forecast of electricity consumption into three categories: short, medium, and long term. The first category refers to the prediction with hourly resolution of the load for a time ranging from one hour to several days. The medium-term forecast relates to a horizon of one to several months ahead. Finally, the long-term forecast is usually associated with periods of one or several years in the future.

To develop the forecast of electrical energy consumption, there are a variety of techniques or algorithms that generate different types of models. This is how in [2] they propose that for the short-term time series analysis models, artificial neural networks, support vector machines, among others, could be used. For the medium term, they propose econometric models, neural networks, linear regression, trend analysis, and end-use models. Therefore, it would be useful to have a methodology that would allow selecting the best forecast model for electrical energy consumption for a particular forecast horizon.

Therefore, the objective of this research is to develop a multi-criteria methodology for the selection of the optimal medium-term forecast model, based on data science. The determination coefficient R^2, the square root of the mean square error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE) are used as decision criteria, obtained from the evaluation of the models, and the last three metrics derived from the forecast with new data. Likewise, the technique of paired comparisons of the hierarchical analytical process (AHP) is used to determine the weights of relative importance of the criteria and the technique of order of preference by similarity to ideal solution (TOPSIS) as the technique of multicriteria decision making. The alternatives to consider in the selection process are the forecast models that are obtained by applying machine learning algorithms.

A bibliographic review was made on the research topic, some of which are mentioned below. For example, in [3], they propose an ANP-TOPSIS multicriteria methodology to select the best forecast model, which they test using a case study of a plastic bag manufacturer and considering forecast models related to time series analysis. The performance metrics of the models are used as decision criteria, and the forecast models to be evaluated are used as decision alternatives. The authors in [4] use the Electre I method to select the set of explanatory variables to be used in the forecast models. The forecast models compared were artificial neural network, arima, autoregressive, exponential smoothing, radial basis function networks, and machine learning. Using the RMSE, MAE and MAPE performance metrics, they conclude that the neural network models had the best performance of all the models. In [5], the authors perform short-term forecasting of individual consumers using four machine learning techniques: support vector machine, artificial neural network, long short-term memory network, and fuzzy adaptive resonance theory neural networks. The metrics used to evaluate the models were R^2 and MAPE. They conclude that the vector support regression model had the best performance of the models considered. In [6] they examined electric power forecasting methods for the short term. They considered classical regression methods and machine learning algorithms. They used forecast accuracy as a performance metric and conclude that all their models had high accuracies. In [7] the authors propose a hybrid electric charge forecast model for the short-medium term composed of principal component analysis and traditional multiple regression, whose results compared classical statistical models. The performance metrics for the comparison of the models were: RMSE, MAE, and MAPE. They conclude that their model is superior to all other models. The authors in [8] propose a framework for feature selection, extraction, and regression to carry out electric charge prediction. They use extreme gradient boosting and random forest to determine the importance of each feature. For the forecast phase they use an improved support vector machine and an improved artificial neural network. For the performance evaluation of the models, they make use of: MAE, MSE, RMSE, and MAPE. The simulation results illustrate that the proposed improved models have higher accuracies than traditional models. In [9] they propose a solution for forecasting the monthly electricity demand using statistical methods (exponential smoothing, arima, and prophet) whose results were compared with other models using the metrics: interquartile range (IQR), median absolute porcentual error (MdAPE), MAPE, and RMSE. They conclude that the combined models perform better than the individualized models. In [10] they review methods for forecasting electricity consumption. From their review, they conclude that linear regression continues to be the most widely used model for the long-term forecast horizon, while neural networks, support vector machines, and fuzzy logic are the most widely used techniques for the short term.

The rest of the article is organized as follows. Section 2 presents the methodology used. Then the results are presented and discussed, to finally present the conclusions derived from the research, and the bibliographical references used.

II. Materials and methods

In this section, the methodology to be used is presented, as well as the data use to illustrate it. Specifically, the methodology consists of the combination of multi-criteria decision-making methods with the methodology of data science projects, incorporating a selection stage of the optimal forecast model.

Multi-criteria decision making (MCDM) resides in addressing a decision problem in which there is more than one decision criteria to consider for the selection of the best alternative. According to [11], MCDM is divided into multi-objective decision making (MODM) and multi-attribute decision making (MADM). The MODM is characterized by having an explicit goal and a continuous decision space (infinitely many alternatives and attributes), while the MADM is characterized by having an implicit goal and a discrete decision space, with discrete alternatives and attributes.

On the other hand, data science is made up of three distinct and overlapping areas: the statistical skill of modeling and summarizing data sets; the computer skills to design and use algorithms that allow this data to be stored, processed, and visualized efficiently; and mastery of the technical area of interest [12]. In general, data science projects consist of a series of stages that can be applied sequentially or with feedback, depending on the case study. In this sense, authors in [13] presents a total of six stages, and in the last one it is incorporated the multicriteria selection of the optimal model.

The first of the stages is the definition of the objective of the data science project, and in this case, it is nothing more than developing the forecast of the consumption of residential electrical energy. Next, there is the stage of extracting or obtaining the data set to be analyzed, which could come from internal or external sources, or a combination of these two. Subsequently, there is the processing of the data, which could include, among others, the detection and imputation of missing data, the detection and imputation of outliers, the detection of duplicate data, verification of the proper format of the data, the transformation of data, and data combination. This stage is the one that usually occupies the most time in projects of this nature. Then there is the exploratory analysis of the data, in which univariate and/or multivariate statistical, analytical, and graphic tools are applied to the processed data. For this research, the modeling stage consists of using machine learning algorithms to generate forecast models of residential electricity consumption. For each of the models obtained, performance metrics are calculated in the training stage of the model, and they are also calculated in the forecast stage, comparing the predicted values with the actual values. The performance metrics to consider are: R^2, MAE, RMSE, and MAPE. Next, the model selection stage is incorporated, in which the paired comparisons technique of the AHP method is used to determine the weights of relative importance of the decision criteria, and the TOPSIS method for the multicriteria selection of the optimal model of forecast.

On the other hand, this section presents the stages of data collection and processing, while the following section presents the stages of exploratory data analysis, data modeling, and optimal model selection.

A. Data collection

The data used to illustrate the methodology was obtained from different sources. This is how the electricity consumption data was extracted from the online platform "Energía Abierta" of the National Energy Commission of Chile [14], which is the regulatory entity of the Chilean energy market. These data correspond to the electricity billed monthly for regulated clients in Chile during the period 2015-2022.

This data set has 390,821 rows and 10 columns. The columns are equivalent to the 10 existing variables, which are: the year in which this billed energy is consumed (“year”), the month in which the billed energy is consumed (“month”), the commune where the distribution company makes the withdrawal of this energy for regulated customers (“comuna”), the type of customers, whether residential or non-residential (“tipo_clientes”), the type of tariff corresponding to the types of customers (“tariff”), the amount of customers that are supplied with the electric power withdrawn from the supply point (“clients”), the base electric power in kWh billed to regulated customers during the reported period (“e1_kwh”), the additional winter electric power in kWh billed to regulated customers (“e2_kwh”), the total electric energy in kWh billed to regulated customers during the reported period (“energy”), and the geographic region in which the billed customers are located (region). Data from the monthly electricity service quality index (“saidi”) were extracted from the same platform. Each one of the 390,821 rows "corresponds to a batch of electrical energy withdrawn from the supply point by the distribution company during the reported period, to supply a certain number of customers, who have the same type of tariff, and who are located in the same region and commune of the country” [15].

The outdoor average temperature data were taken from the website of the General Directorate of Aeronautics of Chile (DGAC) [16] and correspond to the monthly historical series from January 2015 to December 2022 of the average temperature (“temp_med”) measured in the center of the Metropolitan Region of Chile. Finally, the data of the Consumer Protection Index (“IPC”) were obtained from the website of the INE of Chile [17] and corresponds to the previously mentioned monthly historical period.

B. Data processing

The data to be used for the forecast correspond to the metropolitan region of Chile, so the first step consisted in filtering the data to obtain records from only that region. This leaves us with 59,704 records and 9 columns, as the region column is no longer useful. Another filter is then done to achieve a data set of residential customers only, which leaves us with 5,090 records and 8 columns, since the customer type is no longer useful. From January 2015 to December 2022 there are 96 months, so the records corresponding to the same month and year were grouped, to go from 5,090 records to 96 monthly records.

Next, the monthly data set, made up of residential electricity consumption and the number of customers billed, is concatenated with the data on average temperature, the consumer protection index CPI (“IPC”), the CPI percentage rate (“rate_IPC”), and the service quality index, to obtain a data set of 96 monthly records with seven columns corresponding to the variables just mentioned plus the date. Next, the possible existence of missing data and duplicate data is verified, resulting in neither of the two types.

Then, the possible existence of atypical data is checked, of which two were detected for residential electricity consumption, during the months of July and August of the year 2020, just when confinements were implemented due to the pandemic. These outliers were imputed with the average values of the values of the respective month, during the period 2015-2019. Finally, the electrical energy consumption data was adjusted to change its units from kilowatt-hours (kWh) to megawatt-hours (MWh).

III. Discussion and Results

The exploratory analysis seeks to determine patterns in the data, as well as possible significant relationships between the variables. In the modeling of the data, different machine learning algorithms are applied to obtain the forecast models to be evaluated. In the selection of the best model, multicriteria decision-making techniques are applied to obtain the optimal model.

A. Exploratory data analysis

The tools of statistics are then used to explore the data set. First, Table I describes the data using univariate statistics. The average values of the temperature, the CPI, the number of customers billed, and the electrical energy in MWh, are like their median. In the same way, it can be observed that the number of customers billed, and the consumption of electrical energy are the variables that have less variability.

TABLE I
Descriptive Statistics of the Data

Statistic	clients	temp_med	IPC	rate_IPC	saidi	energy
Count	90	90	90	90	90	90
Mean	2,412,983.20	15.62	101.51	0.003	0.76	527,196.73
std	145,229.64	4.74	7.44	0.004	0.61	65,604.23
Min	1,797,917.00	8.20	89.56	-0.004	0.28	423,719.89
Q_1	2,310,053.25	11.53	96.10	0.001	0.45	475,294.21
Median	2,420,15.50	15.55	100.51	0.003	0.57	510,559.67
Q_3	2,523,130.00	20.20	105.11	0.005	0.81	571,258.85
Max	2,797,967.00	24.10	122.48	0.019	5.04	711,243.07

Next, a correlation analysis between the variables is made, using three methods: Pearson, Spearman, and Kendall. The first of the methods is parametric, while the other two are of the non-parametric type [18]. Because the results are similar for the three methods, Fig. 1 presents the correlation matrix, when the Pearson method was used.

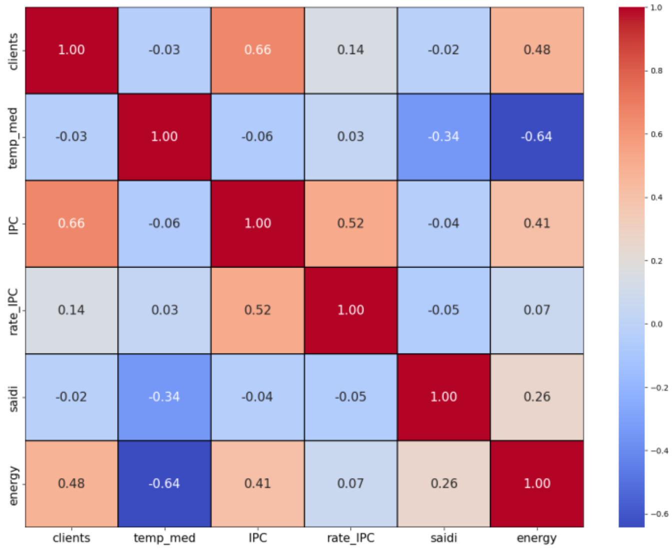

Fig. 1. Correlation matrix.

The correlation coefficient varies between -1 and 1, being negative when the relationship between the variables is inverse, and positive when it is direct. Ratner [19] postulates that “values between 0 and 0.3 (0 and -0.3) indicate a weak positive (negative) relationship. Values between 0.3 and 0.7 (−0.3 and −0.7) indicate a moderate positive (negative) relationship. Values between 0.7 and 1.0 (−0.7 and −1.0) indicate a strong positive (negative) relationship”. In this sense, from Fig. 1 residential electricity consumption (“energy”) has a direct and moderate correlation value of 0.48 with the number of users (clients), inverse and moderate with the ambient temperature (-0.64), direct and moderate with the CPI index (0.41), direct and weak with the saidi service quality index (0.26), and almost zero with the CPI percentage rate.

The inverse relationship between energy consumption and average temperature is confirmed by looking at Fig. 2, which shows monthly average energy consumption vs. the average monthly temperature.

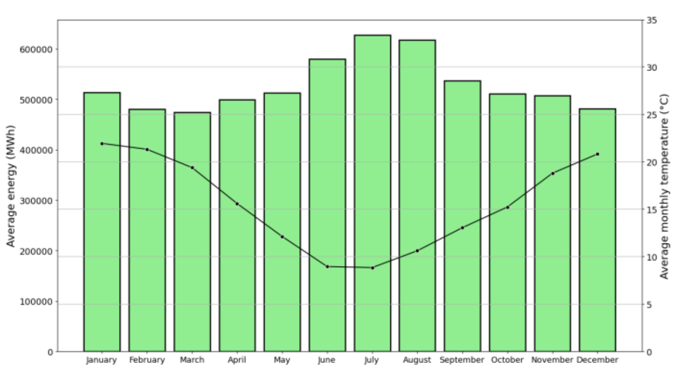

Fig. 2. Residential electricity consumption vs. Average temperature.

It can be noted that during the months of low average temperature, below 10°C during the months of June, July, and August, the consumption of electrical energy reaches its maximum values. Then it reaches its lowest values between the months of December and March.

B. Data modeling

At this point, the algorithms are applied: K nearest neighbors (K-NN), decision tree regression (TDR), support vector regression (SVR), multiple linear regression (MLR), and artificial neural network (ANN), to obtain forecast models, along with their respective performance metrics. In [3], the authors only consider time series analysis models.

The data set to be considered in the application of each of the algorithms corresponds to ninety records, equivalent to ninety months from January 2015 to June 2022. The number of customers billed, the average temperature in degrees Celsius, the consumer protection index, and the saidi service quality index, make up the set of explanatory variables. On the other hand, the consumption of residential electricity billed in megawatt-hours (MWh) represents the objective variable of the models.

Among the performance metrics considered is R^2, which is associated with the quality of fit of the model obtained. It indicates the percentage or proportion of the total variation of the objective variable that is explained by the independent variables of the model [20]. Likewise, the MAE, the RMSE, and the MAPE are used, which are standard statistical measures for forecast models [21].

1) Artificial Neural Network

The first step in the application of ANN algorithm consists of dividing the data into two parts: 70% of the data is used to train the model, while the remaining 30% is used for testing and validation the trained model. The 30% is divided into two equal parts, in such a way that the test and the validation of the model have the same amount of data.

The network model with which we work is known as a "multilayer perceptron" (MLP), which is composed of different layers: an input layer, an output layer, and a set of hidden layers located between the input and the output [22]. In this case, three layers were considered, all the “dense” type, which implies that the corresponding layer connects all its neurons with all the neurons of the preceding layer [23]. The input layer has 256 neurons, the hidden layer also has 256 neurons, and the output layer, which, since the model is regression, only has one neuron. Each one of the layers has an activation function, which is more than a mathematical function that is applied to the heavy sum of the input signals of the neuron, to obtain the output of said neuron. There is a variety of activation functions, among which are: “sigmoid”, “tanh”, “ReLU”, “ELU”, “Leaky ReLU” [22]. Additionally, loss functions must be defined, to control the deviation of the forecast when it is compared with the desired value [24]. In addition, when the deviation is significant, the output is fed back to the input through the optimizer, to update the weights of the inputs and therefore the training cycle is repeated. The iterations through the entire data set of the network training process are known as the “epoch” [25].

In the study network, a ReLU type activation function was defined, both for the input layer and the hidden layer, to limit the values of their respective outputs to only positive values, while the output layer has a "linear" type activation function in order not to limit the prognosis. As loss functions we work with the MAE and the MSE, which are typical in regression problems. Likewise, the root mean square propagation optimizer (RMSProp) is used to update the weights. Finally, for the history of the network a total of 2,000 times is set.

The previously mentioned performance metrics are used to evaluate the model. While, for the analysis of the residues, the Shapiro-Wilk test is used, which according to [26] has as null hypothesis that the data is normally distributed. The test statistic varies between 0 and 1, and when it is close to 1 it is an indication that the data is normally distributed.

After evaluating the model, a relatively good quality of fit of 73.8% was obtained, the MAE obtained was 23,551.24 MWh, which represents around 4.50% of the average of the electric power consumption of the historical data, and the RMSE obtained of 27,920.69 MWh represents 5.30% of the average just mentioned. Likewise, a MAPE of only 4.63% was obtained. On the other hand, from the analysis of the residuals, it can be said that they are normally distributed since the Shapiro-Wilk test statistic is close to unity (0.964) and the p-value is notably greater than 5% of statistical significance (0.818).

2) Multiple Linear Regression

For the application of MLR algorithm, the data is initially divided to assign 70% to the model training process, and the remaining 30% are used for model testing.

After applying the algorithm, the model obtained was evaluated, from which a quality of fit of almost 63% was achieved, lower than that obtained with the previous model, but still an acceptable value. Regarding the RMSE, it was 40,787.96 MWh, which represents 7.75% of the average residential energy consumption of the historical data, the MAE was 33,635.32 MWh, which represents almost 6.40% of the average just mentioned, and for the MAPE a value of 6.35% was obtained. These last three results are superior to those obtained with the ANN model, but they can still be considered low. Regarding the residuals, a p-value significantly greater than 5% of statistical significance (0.53) was obtained, and the test statistic is close to unity (0.967), so it can be said that the residuals are normally distributed.

3) K Nearest neighbors

The K-NN algorithm can be used for both classification and regression, but the basic principle is the same in both ways, i.e., consider elements that are like each other (neighbors). It is a non-parametric technique in the sense that it has no restrictions on the distribution of the data. In the case of regression, the data to be forecast is estimated through a statistic, which is usually the mean value, which is obtained by summarizing the characteristics of the nearest neighbors [27]. For this algorithm, the division of 70% of the data was made for the training of the model, and the remaining 30% for the test of said model.

After using the K-NN regression algorithm, by means of the optimal value 5 for the number of nearest neighbors, the results of the evaluation of the model with the test data were reached. It was obtained that around 67% of the variability of the model is explained by the independent variables, the RMSE of 37,442.99 MWh represents 7.10% of the average value of the data on electric power consumption of the data set, the MAE of 31,090.85 is equivalent to 5.90% of the mean value just mentioned, and the MAPE is only 5.89%. Regarding the analysis of the residuals, it could be said that they are normally distributed since the Shapiro-Wilk test statistic is close to unity (0.978) and the p-value is greater than 5% of statistical significance (0.806).

4) Decision Tree Regression

According to those proposed by [28] "Decision trees form classification and regression models as a tree structure by asking questions and creating decision rules according to the structure of the data sets that constitute a problem." The use of the decision tree in regression problems is a robust alternative to parametric techniques since it does not require assumptions about the data set under study.

For the application of this technique, in the first place, the division of the data set is carried out, assigning 70% for the model training process, and 30% for its evaluation. After the application of the algorithm, the forecast model was obtained whose metrics derived from its evaluation are an acceptable fit quality of the model of 66.4%, the RMSE of 39,368.92 MWh is equivalent to around 7.50% of the average value of electrical energy consumption from the historical data, the MAE of 29,591.82 MWh represents 5.60% of the mean value just mentioned, and the MAPE obtained was 5.72%.

5) Support vector regression

Initially, support vector machines were created to deal with classification problems, but later, adjustments were made so that they could be used in regression problems based on the same operating principle. The support vector algorithm uses the kernel concept to convert the given data into a higher dimension, to achieve the hyperplanes. The points located on each side of the hyperplane and closest to it are known as support vectors. There are four main types of kernels, namely linear, polynomial, sigmoid, and radial basis function [29]. In this investigation the radial basis function kernel is used. Its use is recommended in cases with few records, it also tolerates problems with dimensions that are not so low, but that do not exceed the number of records.

As in the previous cases, the data is divided into two parts, 70% for model training and 30% for model testing. After applying the algorithm, the model was obtained and it was evaluated using the R^2, RMSE, MAE, and MAPE metrics. The model's quality of fit is only 54.8%, the lowest value of all the models considered, although a low R^2 value by itself does not indicate the presence of a “bad” model. The RMSE value of 35,119.12 MWh is equivalent to 6.66% of the average value of the historical data of electricity consumption and the MAE value of 29,772.01 MWh is equivalent to 5.65%, while the MAPE was only 5.50%. Regarding the residuals, it can be said that they are normally distributed since the test statistic is close to unity (0.983) and the p-value is significantly greater than 5% (0.919).

6) Forecast Comparison

As already mentioned, the corresponding model for each of the algorithms was trained with 70% of the historical data, and with the other 30%, performance metrics were generated to assess the forecast quality of the respective model. However, sometimes it is convenient to re-evaluate the respective model, but using data outside the original data set. The authors in [3] do not consider the calculation of the metrics in the forecast with new data.

Then, once each of the models was trained and tested, they were used to generate the forecast for residential electricity consumption for the next six months. Subsequently, this forecast is used, and the data records from the month of July of the year 2022 to the month of December of the same year, to recalculate the RMSE, MAE, and MAPE metrics. The results are presented in Table II. Being metrics of the type "the less the better", it can be deduced that the MLR model is the one that has the best performance in forecasting new data for this objective variable.

TABLE II
MODEL METRICS AFTER FORECAST

Métrica	ANN	MLR	K-NN	TDR	SVR
RMSE (MWh)	147,805.14	39,228.61	59,711.80	54,836.59	101,100.79
MAE (MWh)	125,879.24	30,522.03	54,093.18	49,033.78	95,359.50
MAPE(%)	16.38	4.50	9.12	7.80	17.30

C. Multi-criteria selection of the best forecast model

There are different decision problems, depending on the situation to be solved. In this sense, there are: selection problems, classification problems, hierarchy problems, and description problems. In this research we are interested in the selection problem, which is one that aims to "select the best individual option or reduce the group of options to a subset of equivalent or incomparable 'good' options." [30].

To deal with selection problems, one of the various multicriteria techniques available can be used. For selection problems, the available techniques are varied, but in this research the AHP is used to define the weights of relative importance of the decision criteria, and TOPSIS for the selection of the best forecast model. As stated by [31] the AHP technique is easy to use, scalable, and its hierarchical structure can be adapted to many problems, while the TOPSIS technique has a simple procedure, is easy to use and program, and the number of steps remains the same regardless of the number of criteria.

On the other hand, a multi-attribute decision problem could be represented through its decision matrix [32]. This is a matrix (M×N) in which the element a_{ij} of the matrix indicates the performance of the alternative A_i when it is evaluated in terms of the decision criterion C_j, (for i = 1, 2, 3,…,M, and j = 1, 2, 3, ..., N). Each of the criteria has a relative importance weight w_j, which is generally defined by the “decision maker”. Then, given a set of alternatives and a set of decision criteria (attributes), one seeks to determine the optimal alternative with the highest degree of "desirability" with respect to the decision criteria.

1) Calculation of the relative importance weights of the criteria

The AHP technique is based on what is known as “paired comparisons”, which consists of taking a couple of items and comparing them with respect to one characteristic, without worrying about other characteristics or other elements. To make the comparisons, Saaty proposed a scale of numbers that indicates how many times an element is more important with respect to another element, according to the criterion or property used to compare them. The intensity of importance goes from "1" to "9", being "1" when the elements have equal importance and "9" when the element is extremely more important with respect to the other element [33].

The criteria considered for the decision matrix are the metrics obtained in the evaluation of the models, namely R^2, $RMSE_1$, MAE_1 and $MAPE_1$, plus the metrics obtained from the forecast of residential electricity consumption during the second half of 2022: $RMSE_2$, MAE_2 and $MAPE_2$. In [3] they only take the metrics in the evaluation stage of the respective model as criteria. To these criteria is incorporated the duration time, in seconds, of the respective algorithm run, to obtain and evaluate each one of the models. Table III shows the matrix of paired comparisons obtained, using the numerical scale of importance intensities.

TABLE III
MATRIX OF PAIRED COMPARISONS

Criteria	R^2	$RMSE_1$	MAE_1	$MAPE_1$	$RMSE_2$	MAE_2	$MAPE_2$	Time
R^2	1.00	2.00	2.00	1.00	1.00	1.00	0.50	3.00
$RMSE_1$	0.50	1.00	1.00	0.50	1.00	1.00	0.50	2.00
MAE_1	0.50	1.00	1.00	0.50	1.00	1.00	0.50	2.00
$MAPE_1$	1.00	2.00	2.00	1.00	1.00	1.00	0.50	3.00
$RMSE_2$	1.00	1.00	1.00	1.00	1.00	1.00	0.50	2.00
MAE_2	1.00	1.00	1.00	1.00	1.00	1.00	0.50	2.00
$MAPE_2$	2.00	2.00	2.00	2.00	2.00	2.00	1.00	3.00
Time	0.33	0.50	0.50	0.33	0.50	0.50	0.33	1.00

Once the paired comparisons matrix is available, we proceed to obtain the weights of relative importance of the criteria. To obtain them, a practical procedure is applied that consists of raising the matrix of paired comparisons to a sufficiently large power, adding by rows, and normalizing these values by dividing the sum of each row by the total sum, stopping the process when the difference between two consecutive powers is minimal [34]. Table IV shows the relative importance weights obtained, the forecast MAPE resulted with the highest importance weight value, while the duration time resulted with the lowest weight value. It can also be seen that the sum of the weights is equal to unity. According to [34], once the weights of relative importance have been obtained, the consistency of the decision maker must be evaluated calculating the consistency ratio (CR), an index that is given as the quotient between the index of consistency (CI) and the random consistency index (ICA). If the CR is less than 10%, it can be said that the decision makers have been consistent, and the weights obtained are validated. For the case of eight alternatives, the value of ICA is equal to 1.41, and the CR was 1.28%, so the relative importance weights are valid.

TABLE IV
WEIGHTS

Criteria	Weight
R^2	0.147
$RMSE_1$	0.098

MAE_1	0.098
$MAPE_1$	0.147
$RMSE_2$	0.116
MAE_2	0.116
$MAPE_2$	0.225
Time	0.054

2) Model selection using TOPSIS

The TOPSIS technique is based on selecting the best alternative by measuring the shortest geometric distance to the ideal positive solution, and the longest geometric distance to the ideal negative solution [35]. It consists of a series of steps, the first of which, common to all multicriteria decision-making techniques, consists of obtaining the decision matrix, which is presented in Table V. Subsequently, the normalized decision matrix is obtained, and then the weighted normalized decision matrix, the positive ideal solution and the negative ideal solution, the distance of each alternative from those solutions, and finally the relative closeness of each alternative to the ideal solution [36].

TABLE V
DECISION MATRIX

Model	R^2	$RMSE_1$	MAE_1	$MAPE_1$	$RMSE_2$	MAE_2	$MAPE_2$	Time
ANN	0.74	27,921	23,551	4.63	147,805	125,879	16.38	127
MLR	0.63	40,787	33,635	6.35	39,229	30,522	4.50	0.75
K-NN	0.67	37,442	31,091	5.89	59,712	54,093	9.12	4.53
TDR	0.66	39,368	29,592	5.72	54,837	49,034	7.80	3.05
SVR	0.55	35,119	29,772	5.50	101,101	95,360	17.30	3.51

The normalized decision matrix is obtained by applying (1) to obtain each normalized element r_{ij} from each element a_{ij}. Then the weighted normalized decision matrix is obtained by using (2) and thus getting each weighted element v_{ij} from each r_{ij}. The weighted normalized decision matrix is presented in Table VI.

$$r_{ij} = \frac{a_{ij}}{\sqrt{\sum_{i=1}^{m} a_{ij}^2}} \quad (1)$$

$$v_{ij} = w_j \cdot r_{ij} \quad (2)$$

TABLE VI
WEIGHTED NORMALIZED DECISION MATRIX

Model	R^2	$RMSE_1$	MAE_1	$MAPE_1$	$RMSE_2$	MAE_2	$MAPE_2$	Time
ANN	0.074	0.034	0.035	0.054	0.085	0.083	0.136	0.054
MLR	0.063	0.049	0.050	0.074	0.023	0.020	0.037	0.000
K-NN	0.068	0.045	0.046	0.068	0.035	0.035	0.076	0.002
TDR	0.067	0.047	0.044	0.066	0.032	0.032	0.065	0.001
SVR	0.055	0.042	0.044	0.064	0.058	0.063	0.144	0.001

From the matrix of Table VI, the positive ideal solution A* is obtained by taking the maximum value of R^2, and the minimum value of the rest of the criteria. In the same way, the negative ideal solution A- is obtained by taking the minimum value of R^2, and the maximum value of the rest of the criteria. Results are shown in (3) and (4).

$$A^* = \{0.074, 0.034, 0.035, 0.054, 0.023, 0.020, 0.037, 0.000\} \quad (3)$$

$$A^- = \{0.055, 0.049, 0.050, 0.074, 0.085, 0.083, 0.144, 0.054\} \quad (4)$$

The next step is to calculate the distance of each alternative with the positive ideal solution Di*, and with the negative ideal solution Di-, and from these values obtain the relative closeness of each alternative with the ideal solution Ci*, by using (5). This last parameter varies between 0 and 1, and the optimal alternative is the one with the highest value. The results are presented in Table VII, from which the model obtained with the Multiple Linear Regression algorithm is the optimal alternative for forecasting the monthly consumption of residential electricity. Note that the ANN model, which had the best performance during the model training phase, is ultimately ranked last, according to this technique and decision criteria.

$$C_i^* = \frac{D_i^-}{D_i^- + D_i^*} \quad (5)$$

TABLE VII
RELATIVE CLOSENESS

Model	D_i^*	D_i^-	C_i^*
ANN	0.143	0.036	0.200
MLR	0.031	0.149	0.826
K-NN	0.049	0.111	0.696
TDR	0.038	0.121	0.760
SVR	0.123	0.064	0.342

IV. CONCLUSIONS

It was presented and illustrated a multicriteria methodology to select the best regression model to forecast residential electrical consumption in the medium term. The TOPSIS multicriteria decision-making technique was applied to select the best residential electricity consumption forecast model, with the Multiple Linear Regression model selected according to this technique, and with the decision criteria used.

By applying the technique of paired comparisons of the AHP, it was possible to obtain the weights of relative importance of the decision criteria, resulting in the MAPE of the forecast obtaining the highest weight of relative importance with 0.225. The duration of the run to obtain the models resulted in the least weight of relative importance.

Machine learning algorithms are applied to obtain forecast models for residential electricity consumption, using the number of users, the average temperature, the CPI index, and the saidi service quality index as explanatory variables. According to the performance metrics for the evaluation of the models, R^2, RMSE, MAE and MAPE, the ANN model had the best performance, followed by the K-NN model.

The models obtained were used to forecast energy consumption for the second half of 2022. The models with the previously mentioned error metrics were evaluated again, resulting in the MLR model having the best performance in the forecast phase with new data, followed by the TDR model.

It is recommended to develop a research to select the best features for the forecast of electrical residential consumption through a multicriteria methodology.

REFERENCES

[1] L. L. Grigsby, Power Systems, Boca Raton, FL: CRC Press, 2007.

[2] V. Gupta y S. Pal, "An Overview of Different Types of Load Forecasting Methods and the Factors Affecting the Load Forecasting", International Journal for Research in Applied Science & Engineering Technology, pp. 729-733, 2017.DOI: http://doi.org/10.22214/ijraset.2017.4132.

[3] Y. Badulescu, A.-P. Hameri y N. Cheikhrouhou, "Evaluating demand forecasting models using multi-criteria decision-making approach", Journal of Advances in Management Research, pp. 2021. DOI: https://doi.org/10.1108/JAMR-05-2020-0080

[4] C. Deina, J. L. Ferreira dos Santos, L. H. Biuk, M. Lizot, A. Converti, H. Valadares Siqueira y F. Trojan, "Forecasting Electricity Demand by Neural Networks and Definition of Inputs by Multi-Criteria Analysis", Energies, p. 2023. DOI: https://doi.org/10.3390/en16041712.

[5] J. V. Jales Melo, G. R. Soares Lira, E. G. Costa, A. F. Leite Neto y I. B. Oliveira, "Short-Term Load Forecasting on Individual Consumers", Energies, p., 2022. DOI: https://doi.org/10.3390/en15165856.

[6] S. Filipova-Petrakieva y V. Dochev, "Short-Term Forecasting of Hourly Electricity Power Demand", Engineering, Technology & Applied Science Research - Reggresion and Cluster Methods for Short-Term Prognosis, vol. 12, nº 2, pp. 8374-8381, 2022. DOI: https://doi.org/10.48084/etasr.4787

[7] G. P. Papaioannou, C. Dikaiakos, A. Dramountanis y P. G. Papaioannou, "Analysis and Modeling for Short- to Medium-Term Load Forecasting Using a Hybrid Manifold Learning Principal Component Model and Comparison with Classical Statistical Models (SARIMAX, Exponential Smoothing) and Artificial Intelligence Models (ANN, SVM): Th", MDPI energies, vol. 9, nº 8, 2016. DOI: https://doi.org/10.3390/es9080635

[8] S. R. Khan, I. A. Hayder, S. u. Rehman Khan, M. A. Habib, M. Ahmad, S. M. Mohsin, F. A. Khan y K. Mustafa, "Enhanced Machine-Learning Techniques for Medium-Term and Short-Term Electric-Load Forecasting in Smart Grids", Energies, p, 2022. DOI: https://doi.org/10.3390/en16010276.

[9] P. Pełka, "Analysis and Forecasting of Monthly Electricity Demand Time Series Using Pattern-Based Statistical Methods", Energies, p, 2023. DOI: https://doi.org/10.3390/en16020827.

[10] M. A. Hammad, B. Jereb, B. Rosi y D. Dragan, "Methods and Models for Electric Load Forecasting: A Comprehensive Review", Logistics, Supply Chain, Sustainability and Global Challenges, vol. 11, nº 1, pp. 51-76, 2020. DOI: https://doi.org/10.2478/jlst-2020-0004.

[11] F. Eltarabishi, O. H. Omar, I. Alsyouf y M. Bettayeb, "Multi-Criteria Decision Making Methods And Their Applications– A Literature Review", de Proceedings of the International Conference on Industrial Engineering and Operations Management, Dubai, UAE, 2020.

[12] J. VanderPlas, Python Data Science Handbook - Essential Tools for Working with Data, Sebastopol, CA: O'Reilly Media, Inc., 2017.

[13] D. Cielen, A. D. B. Meysman y M. Ali, Introducing Data Science, Shelter Island, NY: Manning Publications Co., 2016.

[14] Comisión Nacional de Energía de Chile, "Energía Abierta", 01 Abril 2023. [En línea]. Available: http://energiaabierta.cl/categorias-estadistica/electricidad/?sf_paged=2. [Último acceso: 16 July 2022].

[15] C. Yajure Ramírez, "Uso de algoritmos de aprendizaje automático para analizar datos de energía eléctrica facturada. Caso: Chile 2015 – 2021", Revista I+D Tecnológico, pp. 17-31, 2022. DOI: https://doi.org/10.33412/idt.v18.2.3678.

[16] Dirección General de Aeronáutica Civil de Chile, "DGAC CHILE", 01 04 2023. [En línea]. Available: https://climatologia.meteochile.gob.cl/application/informacion/fichaDeEstacion/330020. [Último acceso: 01 12 2022].

[17] Instituto Nacional de Estadísticas de Chile, "INE", 01 04 2023. [En línea]. Available: https://www.ine.gob.cl/estadisticas/economia/indices-de-precio-e-inflacion/indice-de-precios-al-consumidor. [Último acceso: 01 12 2022].

[18] A. Navlani, A. Fandango y I. Idris, Python Data Analysis, Birmingham, UK: Packt Publishing Ltd., 2021.

[19] B. Ratner, Statistical and Machine-Learning Data Mining - Techniques for Better Predictive Modeling and Analysis of Big Data, Boca Raton, FL: CRC Press Taylor & Francis Group, 2017.

[20] D. N. Gujarati y D. C. Porter, Econometría, México, D. F.: McGraw-Hill/Interamericana Editores, S.A. DE C.V., 2010.

[21] S. Makridakis, S. Wheelwright y R. Hyndman, Manual of Forecasting: Methods and Applications, 1997.

[22] A. Kapoor, A. Gulli y S. Pal, Deep Learning with TensorFlow and Keras, Birmingham: Packt Publishing Ltd., 2022.

[23] J. Moolayil, Learn Keras for Deep Neural Networks - A Fast-Track Approach to Modern Deep Learning with Python, Vancouver, BC, Canada: Apress, 2019. DOI: https://doi.org/10.1007/978-1-4842-4240 -7

[24] F. Chollet, Deep Learning with Python, Shelter Island, NY: Manning Publications Co., 2018.

[25] J. Brownlee, Deep Learning With Python - Develop Deep Learning Models On Theano And TensorFlow Using Keras, Melbourne, Australia: Machine Learning Mastery, 2016.

[26] J. Arnastauskaite, T. Ruzgas y M. Braženas, "An Exhaustive Power Comparison of Normality Tests", Mathematics, p. 2021. DOI: https://doi.org/10.3390/math9070788.

[27] M. E. Fenner, Machine Learning with Python for Everyone, Boston: Pearson Education, Inc., 2020.

[28] B. K. Gacar y İ. D. Kocakoç, "Regression Analyses or Decision Trees?", Manisa Celal Bayar University Journal of Social Sciences, pp. 251-260. DOI: https://doi.org/10.18026/cbayarsos.796172

[29] R. Muthukrishnan y M. Jamila. S, "Predictive Modeling Using Support Vector Regression", International Journal of Scientific & Technology Research, pp. 4863-4865, 2020.

[30] A. Ishizaka y P. Nemery, Multi-Criteria Decision Analysis - Methods and Software, West Sussex, United Kingdom: John Wiley & Sons, Ltd, 2013. DOI: https://doi.org/10.1002/9781118644898

[31] M. Velasquez y P. T. Hester, "An Analysis of Multi-Criteria Decision Making Methods", International Journal of Operations Research, pp. 56-66, 2013. http://www.orstw.org.tw/ijor/vol10no2/ijor_vol10_no2_p56_p66.pdf.

[32] E. Triantaphyllou, B. Shu, S. Nieto Sanchez y T. Ray, "Multi-Criteria Decision Making: An Operations Research Approach", Encyclopedia of Electrical and Electronics Engineering, pp. 175-186., 1998.

[33] T. L. Saaty, "Decision making with the analytic hierarchy process", International Journal of Services Sciences, pp. 83-98, 2008. DOI: https://dx.doi.org/10.1504/IJSSCI.2008.017590..

[34] J. M. Moreno Jiménez, "El proceso analítico jerárquico (AHP). Fundamentos, metodología, y aplicaciones.", Revista Electrónica de Comunicaciones y Trabajos de ASEPUMA, pp. 28-77, 2002.

[35] B. Sahoo, R. N. Behera y P. K. Pattnaik, "A Comparative Analysis of Multi-Criteria Decision Making Techniques for Ranking of Attributes for e-Governance in India", International Journal of Advanced Computer Science and Applications, pp. 65-70, 2022. DOI: https://dx.doi.org/10.14569/IJACSA.2022.0130311.

[36] J. Papathanasiou y N. Ploskas, Multiple Criteria Decision Aid - Methods, Examples and Python Implementations, Cham, Switzerland: Springer Nature Switzerland AG, 2018. DOI: https://doi.org/10.1007/978-3-319-91648-4

César Aristóteles Yajure Ramírez, is an Electrical Engineer graduated from the University of Carabobo, Venezuela, in 1998. Subsequently, he received his Master's degree in Operations Research in 2006 at the Central University of Venezuela, becoming a postgraduate professor starting in 2008. He has experience as an engineer, and as a teacher, highlighting his time at the IUT Dr. Federico Rivero Palacio de Caracas, as a professor in the Department of Electricity during the period 2003-2013, as well as at the School of Electrical Engineering and Industrial Automation of the institute DUOC UC of Santiago de Chile, between 2019 and 2023. Between 2012 and 2017 he worked at the Ministry of Electric Power of Venezuela,

first in the electric development area, and then as coordinator of the electric demand estimation area and final use of energy. He has publications in the areas of: Multi-criteria decision making, electricity demand forecasting, performance evaluation of photovoltaic solar plants, data science, among others.
ORCID: https://orcid.org/0000-0002-3813-7606.