Revista Facultad de Ingeniería Universidad de Antioquia
ISSN: 0120-6230
ISSN: 2422-2844
Facultad de Ingeniería, Universidad de Antioquia

Porteiro, Rodrigo; Hernández-Callejo, Luis; Nesmachnow, Sergio
Electricity demand forecasting in industrial and residential facilities using ensemble machine learning
Revista Facultad de Ingeniería Universidad de Antioquia, no. 102, 2022, January-March, pp. 9-25
Facultad de Ingeniería, Universidad de Antioquia

DOI: https://doi.org/10.17533/udea.redin.20200584

Revista Facultad de Ingeniería, Universidad de Antioquia, No.102, pp. 9-25, Jan-Mar 2022

Electricity demand forecasting in industrial and residential facilities using ensemble machine learning

Predicción de demanda eléctrica en instalaciones industriales y residenciales utilizando aprendizaje automático combinado

Rodrigo Porteiro [1*], **Luis Hernández-Callejo** [2], **Sergio Nesmachnow** [1]

[1]Universidad de la República. Av. 18 de Julio 1824-1850. C. P. 11200. Montevideo, Uruguay.
[2]Departamento de Ingeniería Agrícola y Forestal, Universidad de Valladolid. Campus Universitario Duques de Soria. C. P. 42004. Soria, España.

CITE THIS ARTICLE AS:

R. Porteiro, L. Hernández-Callejo and S. Nesmachnow. "Electricity demand forecasting in industrial and residential facilities using ensemble machine learning", *Revista Facultad de Ingeniería Universidad de Antioquia*, no. 102, pp. 9-25, Jan-Mar 2022. [Online]. Available: https://www.doi.org/10.17533/udea.redin.20200584

ARTICLE INFO:

Received: February 27, 2019
Accepted: May 21, 2019
Available online: May 21, 2019

KEYWORDS:

Energy; forecasting; artificial intelligence

Energía; pronóstico; inteligencia artificial

ABSTRACT: This article presents electricity demand forecasting models for industrial and residential facilities, developed using ensemble machine learning strategies. Short term electricity demand forecasting is beneficial for both consumers and suppliers, as it allows improving energy efficiency policies and the rational use of resources. Computational intelligence models are developed for day-ahead electricity demand forecasting. An ensemble strategy is applied to build the day-ahead forecasting model based on several one-hour models. Three steps of data preprocessing are carried out, including treating missing values, removing outliers, and standardization. Feature extraction is performed to reduce overfitting, reducing the training time and improving the accuracy. The best model is optimized using grid search strategies on hyperparameter space. Then, an ensemble of 24 instances is generated to build the complete day-ahead forecasting model. Considering the computational complexity of the applied techniques, they are developed and evaluated on the National Supercomputing Center (Cluster-UY), Uruguay. Three different real data sets are used for evaluation: an industrial park in Burgos (Spain), the total electricity demand for Uruguay, and demand from a distribution substation in Montevideo (Uruguay). Standard performance metrics are applied to evaluate the proposed models. The main results indicate that the best day ahead model based on ExtraTreesRegressor has a mean absolute percentage error of 2.55% on industrial data, 5.17% on total consumption data and 9.09% on substation data.

RESUMEN: Este artículo presenta modelos de pronóstico de demanda eléctrica industriales y residencial, aplicando aprendizaje automático combinado. El pronóstico de demanda eléctrica a corto plazo beneficia a consumidores y proveedores, ya que permite mejorar las políticas de eficiencia energética y el uso racional de los recursos. Se desarrollan modelos de inteligencia computacional para el pronóstico diario de demanda eléctrica y una estrategia híbrida para construir el modelo de pronóstico diario basado en modelos para la próxima hora. Se aplican tres métodos de preprocesamiento de datos: tratamiento de valores perdidos, eliminación de valores atípicos y estandarización. Se aplica extracción de características para reducir el sobreajuste y el tiempo de entrenamiento, mejorando la precisión. El mejor modelo se optimiza mediante búsqueda de grilla en el espacio de hiperparámetros. Luego se genera un conjunto de 24 instancias para construir el modelo de pronóstico completo para el día siguiente. Las técnicas aplicadas se desarrollan y evalúan en el Centro Nacional de Supercomputación (Cluster-UY), Uruguay. Se utilizan tres conjuntos de datos reales para la evaluación: un parque industrial en Burgos (España), la demanda eléctrica total de Uruguay y la demanda de una subestación de distribución en Montevideo (Uruguay). Se aplican métricas estándar para evaluar los modelos propuestos. Los resultados indican que el mejor modelo, basado en ExtraTreesRegressor, tiene un error porcentual medio de $2,55\%$ en datos industriales, $5,17\%$ en consumo total y $9,09\%$ en subestación.

* Corresponding author: Rodrigo Porteiro
E-mail: rporteiro@ute.com.uy
ISSN 0120-6230
e-ISSN 2422-2844

1. Introduction

Uncertainty is a specific characteristic of the energy sector. Although decisions in the energy sector are generally not

based on predictable outcomes, some variables that affect decision making can be predicted, with certain degree of confidence, using information from different sources [1, 2].

Examples of useful information for decision making is that related to natural variables (temperature, wind speed, etc.). Information related to energy consumption and demand profiles of users is valuable too. Furthermore, new sources of renewable energy generation developed in the last 30 years are directly related to natural variables, and the corresponding information is often incorporated in prediction models for decision making [3].

Due to the aforementioned reasons, a large number of stochastic variables must be taken into account to improve operational decision making, but also to assure that the derived actions are feasible from an economic point of view. When considering a large number of variables, the complexity of the underlying models notoriously increases. However, the increase in complexity associated with the number of variables is partly compensated because the hardware infrastructure to perform computations on large volumes of data has developed strongly.

New challenges have emerged from the described reality. A very relevant one is related to the development of an intelligent system to take advantage of new sources of information and available data. Classic statistical models, that were useful for making predictions some decades ago, have limitations in this new context. Computational intelligence methods have demonstrated excellent forecasting accuracy in different areas, in recent years [4–6]. These methods are robust and tolerant to uncertainty, and they are able to learn the most relevant features of the considered data to provide a precise forecast, thus providing excellent results by excluding non-relevant information and focusing on the most useful data.

In this line of work, this article presents the application of several prediction algorithms based on computational intelligence to forecast the electricity demand of an industrial park in Spain, the electricity demand of a substation in Uruguay and the total electricity demand of Uruguay. The modeled scenarios are based on historical demand data of the industrial park (from 2014 to 2017), historical demand data of the substation in Uruguay (from 2017 to 2018) and historical total demand data of Uruguay (from 2010 to 2018). For the industrial park, a forecasting model for the next 24 hours is built by optimizing the algorithm that presented the best results for the one hour forecast.

Overall, the major contributions of the research reported in this article are: *i*) the evaluation and comparison of computational intelligence models applied to forecasting the demand of an industrial park in Spain, the demand of a substation in Uruguay and the total demand of Uruguay. Also *ii*) the optimization of the proposed models using the high performance computing infrastructure of the National Supercomputing Center, in Uruguay.

This work extends our previous article *Short term load forecasting of industrial electricity using machine learning* [7], presented at II Ibero-American Congress on Smart Cities, Soria, Spain, 2019. The main contributions of the extended version are: i) a residential demand forecasting scenario applying the studied techniques on a substation and incorporating climate variables and ii) a total demand forecasting scenario of Uruguay including residential and industrial consumers. Both new instances are analyzed in order to evaluate different consumer profiles as well as different types of influence of weather variables.

The article is organized as follows. Section 2 presents the formulation of the electricity demand forecasting problem and a review of related works. Section 3 describes the proposed approach to solve the proposed problem. Section 4 reports the experimental analysis of the studied methods, and Section 5 reports analysis of the best method and extension to 24-hour demand forecast is presented. Finally, Section 6 formulates the conclusions and main lines for future work.

2. Energy demand forecasting

This section introduces the energy demand forecasting problem, describes forecasting techniques, and reviews related works.

2.1 General considerations

The energy demand forecasting problem is usually solved applying mathematical methods using historical data for prediction. There is no a general method that can be used in all types of energy demand forecasting. Thus, an appropriate method must be found for each demand profile. Using historical data of a particular demand profile is common in practice to determine the most effective algorithm. The problem can be classified by the time horizon to forecast: ultra short-term demand forecasting (up to a few minutes ahead), short-term demand forecasting (up to few days ahead), medium-term demand forecasting (up to few month ahead), and long-term demand forecasting (years ahead). Different techniques are applied when considering each time horizon. This work focuses on short-term demand forecasting using historical data.

The energy management and operation of electric grids becomes highly difficult and uncertain, particularly when new technologies are incorporated. The power demand of end customers is versatile and changes on hourly, daily, weekly, and seasonally basis. Hence, there is a real need of developing models for accurately forecasting at different time horizons, depending on the management goals.

This work focuses on both industrial and residential power consumption. Residential power profiles are usually variable, mainly depend on the time of the day and the day of the week, but they also depend on occasional vacations and other factors [8]. On the other hand, industrial power profiles tend to be stable, due to the needs of industrial processes themselves.

There are two classes of forecasting models for predicting power demand profiles: statistical and physical models. The main goal of both classes is to predict the power profile at a future time frame. Statistical models can be built for time series analysis.

They are less complex than physical models and are suitable for short term prediction. Physical models are based on differential equations for relating the dynamics of the environment and generally are applied for long term forecasting. In this article, statistical models are selected for short term forecasting due to their very good prediction accuracy and lower complexity.

2.2 Problem formulation and strategies

This section describes the problem formulation and the studied strategies for electric demand forecasting.

Relation between one hour and 24 hour forecasting. This article focuses on applying computational intelligence methods to develop a model for forecasting electricity demand 24 hours ahead. When historical data are available with hourly frequency, is natural to develop a model that predicts next hour. From that model, a multi-step forecasting model can be constructed (i.e., 24 steps in the future).

Four strategies are typically applied for multi-step forecasting starting from a one-step model:

- *Direct strategies* develop a different model for each time step to be predicted. Assuming past observations of the variable to be predicted are used, this strategy implies, in case of 24 steps, developing 24 models with the structure defined in Equation 1, where $pred_t$ is the prediction of time t value and obs_t is the observed value at time t.

$$\begin{aligned} pred_{(t+1)} &= model_1(obs_t, obs_{(t-1)}, ..., obs_{(t-n)}) \\ pred_{(t+2)} &= model_2(obs_t, obs_{(t-1)}, ..., obs_{(t-n)}) \\ &\dots \\ pred_{(t+24)} &= model_{24}(obs_t, obs_{(t-1)}, ..., obs_{(t-n)}) \end{aligned} \tag{1}$$

  Unfortunately, a direct strategy implies developing a model for each time step to be predicted and consequently is very expensive computationally. In addition, temporary dependencies are not explicitly preserved between consecutive time steps.

- *Recursive strategies* apply a one-step model (recursively), multiple times. Predictions for previous time steps are used as input for the prediction on the following time step. The structure to develop for a recursive strategy is presented in Equation 2.

$$\begin{aligned} pred_{(t+1)} &= model_1(obs_t, obs_{(t-1)}, ..., obs_{(t-n)}) \\ pred_{(t+2)} &= model_1(pred_{(t+1)}, obs_t, obs_{(t-1)}, ..., \\ & obs_{(t-n+1)}) \\ &\dots \\ pred_{(t+24)} &= model_1(pred_{(t+23)}, pred_{(t+22)}, ..., \\ & pred_{(t+1)}, obs_{(t-n+23)}) \end{aligned} \tag{2}$$

  In this strategy predictions are used instead of observations. A single model is trained, but the recursive structure allows prediction of errors to accumulate; also, the performance of the model can quickly degrade as the time horizon increases.

- *Hybrid strategies* combine the previously described to get benefits form both methods. A separate model is constructed for each time step to be predicted. Each model may use the predictions made by models at prior time steps as input values. For example, using all known prediction, a hybrid strategy produces the structure in Equation 3.

$$\begin{aligned} pred_{(t+1)} &= model_1(obs_t, obs_{(t-1)}, ..., \\ & obs_{(t-n)}) \\ pred_{(t+2)} &= model_1(pred_{(t+1)}, obs_t, ..., \\ & obs_{(t-n)}) \\ &\vdots \\ pred_{(t+24)} &= model_1(pred_{(t+23)}, pred_{(t+22)}, ..., \\ & obs_t, ..., obs(t-n)) \end{aligned} \tag{3}$$

- *Multiple output strategies* develop a model that has as output all time steps to be predicted (in this case 24). Multiple output models are more complex as they can learn the dependence structure between inputs and outputs as well as between outputs. For this reason,

they are slower to train and require more data to avoid overfitting. Equation 4 shows the corresponding structure.

$$pred_{(t+1,\ldots,t+24)} = model_1(obs(t), obs(t-1), \ldots, obs(t-n)) \quad (4)$$

In this work, hybrid strategies are applied for solving the forecasting problem.

One hour forecasting model training. Section 2.3 reviews different approaches and methods for short term demand forecasting. This work explores the use of machine learning techniques, mainly those based on model ensembles. Feature selection is commonly applied in this kind of problems due to several reasons. Simpler models are easier to interpret, and have shorter training times. Also, the size of the model using less features is smaller, mitigating the *curse of dimensionality* [9]. But the main reason to apply feature selection is to reduce overfitting, enhancing generalization of the model to unseen data.

Once established the strategy to extend the next hour forecasting models to twenty four hours model, the main issue is to obtain the best possible model for the next hour. With this purpose, standard steps are taken: i) data gathering, ii) data preparation, iii) choosing a model, iv) training, v) evaluation, vi) parameter tuning, and vii) testing. Each of these steps is described in detail in Section 3.

Complete model. After obtaining a one-hour model with optimized parameters, it is trained for the next hour taking all steps mentioned. Thus, 24 four different instances of this model are trained, one for each of the next 24 hours. Then, the hybrid strategy described in Equation 3 is applied to build a 24-hour forecasting model. The complete model is evaluated on testing data and results are reported.

2.3 Related works

Several methods have been proposed for electricity demand forecasting, applying short, medium and long-term predictions. These methods are classified in statistical models and machine learning models. This work focuses on short-term demand forecasting using machine learning.

Most used forecasting techniques include auto regressive models (AR), moving average models (MA), auto regressive moving average models (ARMA) and auto regressive integrated moving average (ARIMA) models [10]. These kind of models are easy to implement. ARIMA models for short-term demand forecasting [11] were initially proposed by Hagan and Behr.

Taylor and McSharry compared different ARIMA implementations using load data from multiple countries [12]. Dudek proposed applying a linear regression technique [13]. However, linear models are inadequate to represent the non-linear behavior of electricity demand series and fail to predict the accurate future demand values. Thus, their forecasting accuracy tends to be poor. Some studies try to overcome the aforementioned difficulties considering nonlinear components, obtaining good accuracy metrics [14].

Several studies have been conducted on short-term demand forecasting using non-linear models. For example, Do et al. described a model for predicting hourly electricity demand considering temperature, industrial production levels, daylight hours, day of the week, and month of the year to forecast electricity consumption [15]. Results suggested that consumption is better modeled considering each hour separately. In our work, this strategy is developed and applied. Son and Kim proposed a method based on support vector regression preceded by feature selection for the short-term forecasting of electricity demand for the residential sector. For feature selection, twenty influential variables were considered and the quality of the model improved substantially [16].

Other mid-term demand forecasting studies consider variables such as GDP and prove that are highly correlated with the demand [17]. Peak demand estimation is also crucial to determine future demand, in order to assist future investment decisions [18]. In this article, the decision to consider ensemble models was taken based in the work presented by Burger and Moura, who applied a gated ensemble learning method for short-term electricity demand forecasting and showed that the combination of multiple models yielded better results than the use of a single model [19]. Silva presented a complex feature engineering to build gradient boosted decision trees and linear regression models for wind forecasting; in our work several similar ideas were developed for demand forecasting [20]. De Felice et al. applied several separate models for each hourly period. Each of those models measure variations in electricity demand based on multiple variables [21]. Recent studies in traffic prediction in the context of Internet of Things have shown promising results using advanced artificial intelligence techniques related to those applied in our work [22, 23]. Computational intelligence has also been applied to forecasting and dissagregation of residential energy consumption [8, 24].

The analysis of the related works allowed to conclude that two main issues impact on the forecasting capabilities and

the results quality: the model itself and other preparation and pre-processing techniques.

Several works applied techniques like data normalization, filtering of outliers, clustering of data or decomposition by transformations [25–28] to improve prediction results.

In our research, several data preparation techniques are applied for building a robust approach for short term energy utilization forecasting. Next section describes the proposed approach.

3. The proposed approach for day ahead demand forecasting

This section describes the proposed approach to solve the day-ahead electricity demand forecasting for an industrial park in Spain, a substation in Uruguay, and for the total demand of Uruguay applying the strategies described in Section 2.2. In addition, implementation details of the proposed models are presented.

3.1 General approach

This subsection describes the data and the proposed methodology for electricity demand forecasting.

Data description, preparation, and metrics

Data description. Data for the three studied scenarios are described next. *Industrial park in Burgos, Spain.* The first scenario reported in this article considers historical hourly energy consumption data from an industrial park in Spain. Data were collected between January 2014 and December 2017. The dataset consists of industrial energy consumption measurements. Each measurement is composed of the following fields:

- *Year* (integer), representing the year on which the measure was taken.
- *Month* (integer), indicating the month on which the measure was taken.
- *Day* (integer), indicating the day on which the measure was taken.
- *Hour* (integer), indicating the hour on which the measure was taken.
- *Dayofweek* (integer), indicating the day on which the measure was taken.
- *Workingday* (boolean), indicating whether the measure was taken in a working day or not.
- *Useful* (boolean), indicating whether the measure is valid.
- *Demand* (float), indicating the real power measured.

Substation SB1872 in Montevideo, Uruguay. The second scenario studied in this article considers historical hourly energy consumption data from a substation in *Tres Cruces* neighborhood in Montevideo, Uruguay. *Tres Cruces* is a neighborhood located near the centre of Montevideo that serves 390 citizens distributed in 117 homes with medium socio-economic level [29]. The studied dataset contains residential energy consumption measurements collected between January 2017 and December 2018. Each measurement is composed of the following fields:

- *Year* (integer), representing the year on which the measure was taken.
- *Month* (integer), indicating the month on which the measure was taken.
- *Day* (integer), indicating the day on which the measure was taken.
- *Hour* (integer), indicating the hour on which the measure was taken.
- *Dayofweek* (integer), indicating the day on which the measure was taken.
- *Workingday* (boolean), indicating whether the measure was taken in a working day or not.
- *Useful* (boolean), indicating whether the measure is valid.
- *Temperature* (float), indicating the temperature.
- *Humidity* (float), indicating humidity.
- *Wind speed* (float), indicating the average wind of a specific hour.
- *Demand* (float), indicating the real power measured.

Total demand of Uruguay. The third scenario studied in this article considers the historical hourly energy total demand from Uruguay for a total period of nine years (data were collected between January 2010 and December 2018). Each measurement is composed of the following fields:

- *Year* (integer), representing the year on which the measure was taken.
- *Month* (integer), indicating the month on which the measure was taken.
- *Day* (integer), indicating the day on which the measure was taken.

- *Hour* (integer), indicating the hour on which the measure was taken.
- *Dayofweek* (integer), indicating the day on which the measure was taken.
- *Workingday* (boolean), indicating whether the measure was taken in a working day or not.
- *Useful* (boolean), indicating whether the measure is valid.
- *Temperature* (float), indicating the temperature.
- *Humidity* (float), indicating humidity.
- *Wind speed* (float), indicating the average wind of that hour.
- *Demand* (float), indicating the real power measured.

Data preparation. For the three studied scenarios, data preparation consists in eliminating useless measurements and replacing outliers. A few useless measurements were found (less than 0.0001%) in each dataset, and none of them corresponded to consecutive hours. Thus, they were replaced by the average measure of the previous and next hour. A measurement is considered an outlier when it deviates from the mean by more than three times the standard deviation [30].

Outliers were replaced by the value of the mean, adding or subtracting three times the standard deviation, depending on whether the outlier is higher or smaller than the mean.

Feature standardization was applied to the three scenarios data to avoid typical scale issues. For instance, if a feature in the dataset has a different order of magnitude compared to others then in algorithms where a metric is involved this big scaled feature becomes dominating and needs to be standardized [31].

Finally, from in each of the scenarios, new features were generated from datasets associated with past demand measures to train the models. In particular, the last 48 measures were considered for each record to capture at least two days of consumption pattern directly in the features.

Several visualization analysis were performed to gain an intuitive insight of the information contained in each feature. The most relevant fact confirmed in this preliminary analysis was the daily periodicity of the demand value in all scenarios. The diagram shown in Figure 1 reports the high correlation between actual demand and the demand of the same hour of two days before in the case of the industrial park scenario. Weather data was not considered in the industrial scenario because of the very low correlation. However, in both scenarios that contain residential consumption data, weather variables are highly correlated with demand. Figure 2 presents the relation between temperature and total consumption for the scenario of total demand of Uruguay showing that demand values are higher in extreme temperatures. Data preprocessing was performed using pandas library [32]. A linear regression model M_{sim} was trained using the sklearn toolkit [33], configured with default parameters as benchmark model. New training and test datasets were produced keeping only the relevant features, according to the analysis performed to determine the relative importance of each feature.

Metrics. Three standard metrics were used for evaluation: Mean absolute percentage error (*MAPE*, Equation 5), root mean square error (*RMSE*, Equation 6) and mean absolute error (*MAE*, Equation 7); $real_i$ represents the measured value for $t = i$, $pred_i$ represents the predicted value and n represents the predicted horizon length.

$$MAPE = 100 \times \frac{\sum_{i=1}^{n} \left| \frac{real_i - pred_i}{real_i} \right|}{n} \tag{5}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (real_i - pred_i)^2}{n}} \tag{6}$$

$$MAE = \frac{\sum_{i=1}^{n} |real_i - pred_i|}{n} \tag{7}$$

Training the one hour ahead forecasting models

Once all data were prepared for model training, a four-step procedure was applied for training and evaluation in all the scenarios studied. The four steps are:

1. Training and test sets were generated in a 3:1 proportion. In the industrial park scenario, the training set considered data from 2014 to 2016 and the test set considered data from 2017. In the substation scenario, the training set considered data from January 2017 to June 2018 and the test set from July 2018 to December 2018. In the total demand scenario, the training set uses data from 2010 to 2016 and the test set considered data from 2017 to 2018.
2. A simple base model was trained for benchmarking. Using the trained model, a recursive feature elimination process was performed. The ten most important features are preserved.
3. Several models were trained and compared with the benchmark model.
4. The best model according to *MAPE* metric was chosen.

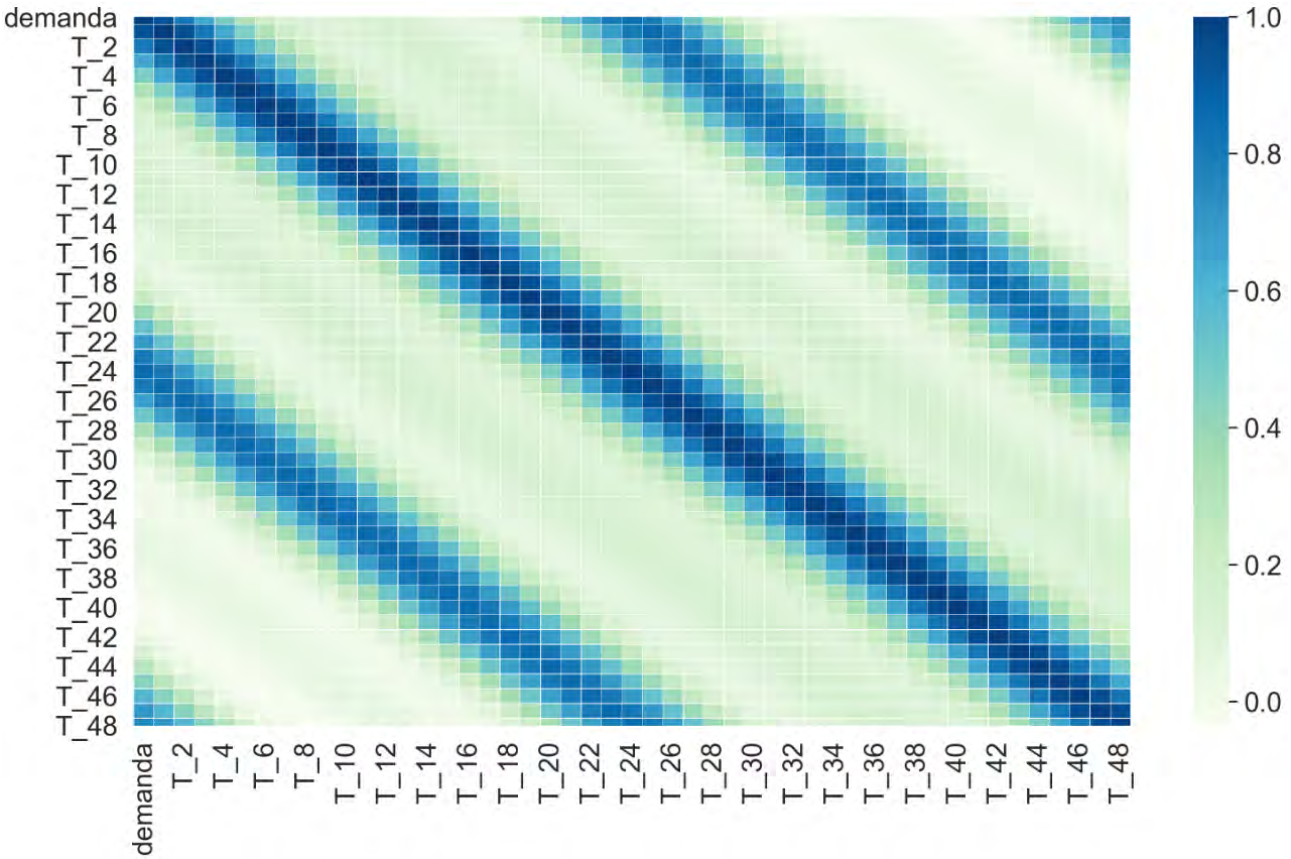

Figure 1 Correlation diagram between actual demand and 48 last demand measures

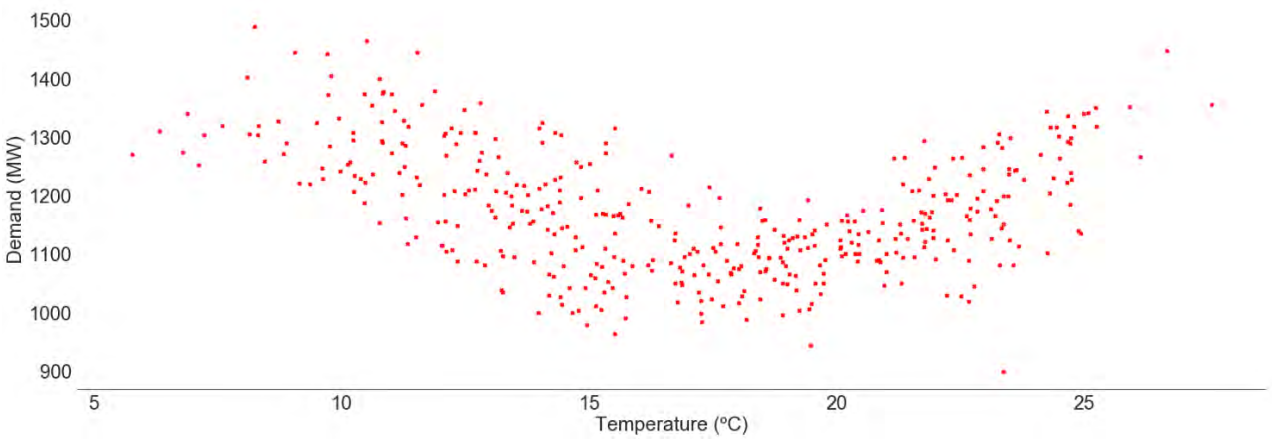

Figure 2 Relation between temperature and total demand of Uruguay

5. An optimization of hyperparameters of the best model was performed using grid search techniques.

Finally, the best model found with the optimized hyperparameters was used as a reference to train the 24 hour forecasting model.

Generation of the 24 hour model

The best model configured with the best hyperparameters obtained in the previous step, was used to generate twenty four models $M_1, M_2, ..., M_{24}$ to forecast day ahead hours. The twenty four models were generated by applying the following procedure:

1. Training and test sets were generated using the same procedure described in Section 3.1. In the industrial

park scenario, the training set considered data from 2014 to 2016 and the test set considered data from 2017. In the substation scenario, the training set considered data from January 2017 to June 2018 and the test set from July 2018 to December 2018. In the total demand scenario, the training set uses data from 2010 to 2016 and the test set considered data from 2017 to 2018.

2. Model M_i was trained using y_i as output, where y_i consists of the demand value corresponding to i hours ahead, and input X is enriched for models M_i, $i > 2$ with a new column consisting of the $i-1$ prediction obtained by the trained model M_{i-1}
3. Models M_i are assembled to get a complete model M to forecast the next 24 hours altogether.

3.2 Implementation

This section describes the implementation of the approach described in Section 3.1.

Computational platform and software

The experimental evaluation was performed in an HP ProLiant DL380 G9 high end server with two Intel Xeon Gold 6138 processors (20 cores each) and 128 GB RAM, from the high performance computing infrastructure of National Supercomputing Center, Uruguay (Cluster-UY) [34].

The proposed approach was implemented in Python. Several scientific packages were used to handle data, train models and visualize results. Used packages included pandas, sklearn, and keras.

A generic module was implemented to train various type of models following a pipeline processing.

Parameter tuning of the studied models were performed using RandomizedSearchCV and GridSearchCV modules from sklearn. The main details of the implementation of the studied models are provided in the following subsections.

Implementation of one-hour model

All one hour models described in this section use training and test sets and data preprocessing presented in Section 3.1.

Base model: Linear regression. A linear regression model was trained to be used as a baseline for the results comparison. A recursive feature selection strategy [35] was also applied on this model for each of the three scenarios to determine the most important features. The rest of features were removed from the dataset.

Ten features were selected based on their relative importance in the industrial demand scenario:

- T_1, T_2, T_{24}, T_{25}: demand values lagged.
- *workingday*: flag indicating whether the day of measured value is a working day
- *month*: month on which the measure was taken.
- *hour*: hour of the day on which the measure was taken.
- *dayofweek*: day of the week on which the measure was taken.
- *day*: day of month on which the measure was taken.
- *year*: year on which the measure was taken.

For the two residential scenarios, the ten most important features were:

- T_1, T_2, T_{24}, T_{25}: demand values lagged.
- *temperature forecast*: temperature external forecast for the hour to be considered.
- *workingday*: flag indicating whether the day of measured value is a working day
- *month*: month on which the measure was taken.
- *hour*: hour of the day on which the measure was taken.
- *dayofweek*: day of the week on which the measure was taken.
- *year*: year on which the measure was taken.

The most relevant past demand values are T_1, T_2, T_{24}, and T_{25} because the current demand is highly correlated with the immediate past demands and also with the demands of the previous day at the same time, due to the daily periodicity. It is worth noting that temperature is the fifth most important feature in scenarios that involve residential demand, in spite of being excluded of the industrial model due to the very low correlation with demand in that case. When training hourly models, temperature external forecast is considered for the corresponding hour.

The full analysis of feature selection experiments is presented and discussed in Section 4.1.

Selection of the best method. Seven regression models were trained for each scenario, including the base model considering the ten most important features and default parameters. The studied models included trained using the scikit-learn API [36]:

Linear Regression, MLP, Extra Trees, Gradient Boosting, Random Forest, K-Neighobors and Ridge.

These models were evaluated using the *MAPE* metric and the linear regression model was used to determine a baseline performance value. The most accurate method was chosen for further evaluation (this method is called M_{best}).

$$\begin{aligned} &pred_{(t+1)} = M_{opt,(t+1)}(obs_t, obs_{(t-1)}, ..., \\ &obs_{(t-n)}) \\ &pred_{(t+2)} = M_{opt,(t+2)}(M_{(t+1)}, obs_t, ..., \\ &obs_{(t-n)}) \\ &... \\ &pred_{(t+24)} = M_{opt,(t+24)}(M_{(t+23)}, M_{(t+22)}, ..., obs_t, ..., \\ &obs(t-n)) \end{aligned} \tag{8}$$

Optimization of the best method. Parameter search techniques were applied for each scenario to optimize a model based on the best method obtained (M_{best}). The model M_{best} trained with default parameters was optimized using two standard tools available in `scikit-learn`:

- GridSearchCV: combines an estimator with a grid search preamble to tune hyper-parameters. The method picks the optimal parameter from the grid search and uses it with the estimator selected according to a predetermined metric.
- RandomizedSearchCV: sets up a grid of hyperparameter values and selects random combinations to train the model and score. After that, the method finds the best parameters setting according to a predetermined metric.

The best parameter set obtained for M_{best} are used in an optimal model M_{opt}. The main details of the implementation of the complete model are described in the next subsection.

3.3 Implementation of the complete model

Model M_{opt} was optimized for predicting the next hour and used for predicting any of the following 24 hours to build the complete model. This decision was adopted assuming that the forecasting quality of the parameter setting obtained in the previous phase is independent of the hour used as output.

To build the complete model, 24 instances of the optimized model M_{opt} were trained. These instances are called $M_{opt,i}$, defining the model trained to forecast the i_{th} hour ahead. The output y_i used to train the model consisted of the demand value for the i-th hour ahead.

For $i > 2$, the input X_i is enriched with a new set of columns consisting in all predictions obtained by models $M_{opt,1}, ..., M_{opt,i-1}$. Then, the complete solution uses a different model for each time step to predict. Predictions for previous time steps are used as input for the prediction on the following time step.

This way, a hybrid strategy is applied to M_{opt}, described in Equation 8.

Finally, the complete model M_{opt} is computed by Equation 9. The output of the model is a 24 valued vector, one prediction for each hour.

$$M_{opt}(t) = (pred_{(t+1)}, pred_{(t+2)}, ..., pred_{(t+24)}) \tag{9}$$

4. Experimental analysis

This section presents the results of the experimental analysis of the proposed computational intelligence methods for day ahead electricity demand forecasting in industrial and residential scenarios.

4.1 Recursive feature elimination

A feature selection analysis was performed using the recursive feature elimination tool available in `sklearn`.

A model is specified and a number of features are selected, and the tool works by recursively removing features and building a new model (of the specified type) on those remaining features.

The accuracy of the new model is used to identify the features or combination of features that contribute the most to predicting the target attribute.

The recursive feature selection tool was applied over the linear regression method described in Subsection 3.2 in each of the three scenarios, to study up to ten features.

Figure 3, 4 and 5 summarize the main results of the analysis, reporting the relative importance of the ten most important features for each scenario.

The most relevant conclusion of the feature selection analysis is the high relative importance of temperature in

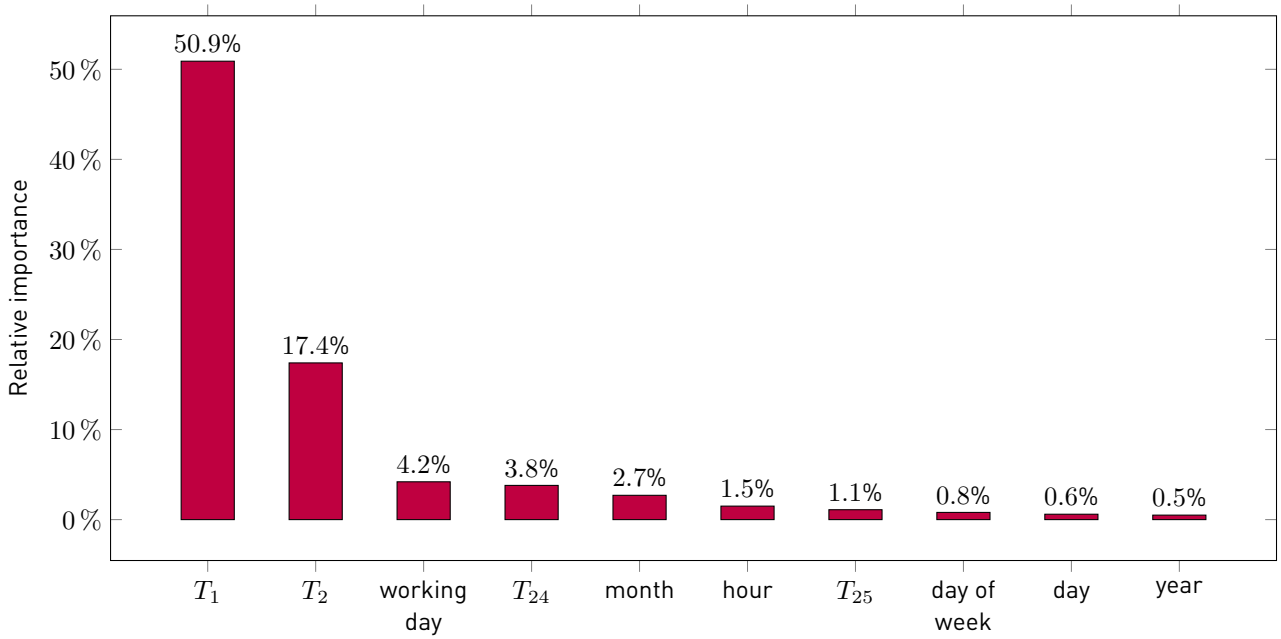

Figure 3 Relative importance of most important features (percentage values), industrial scenario

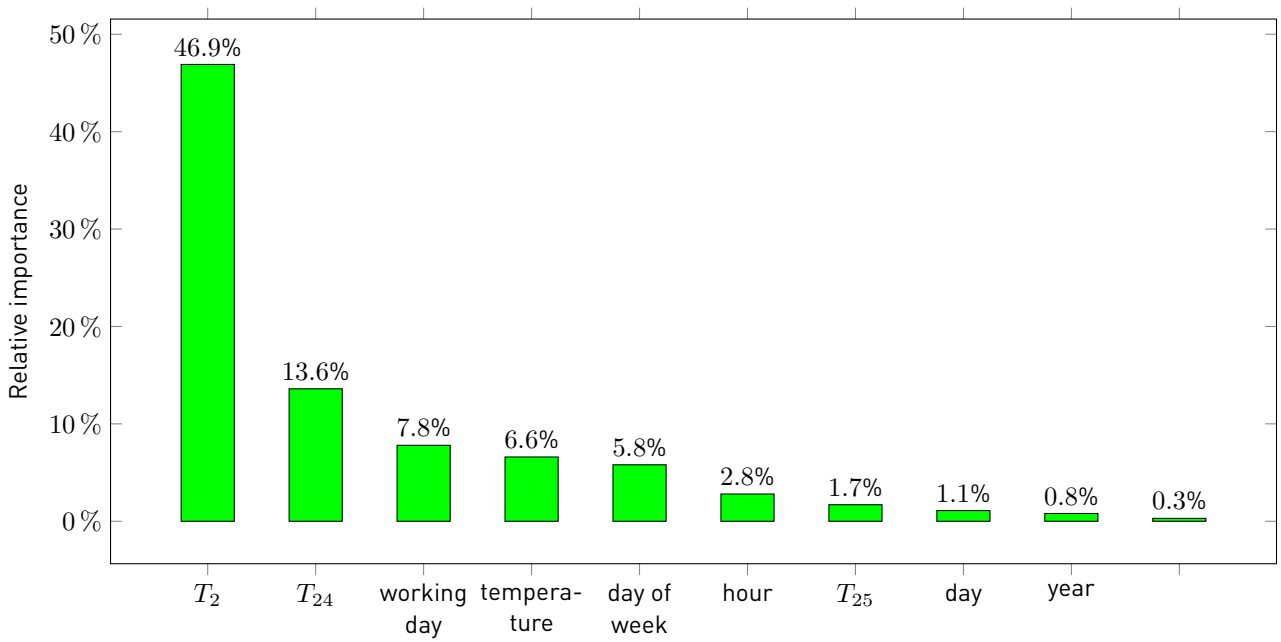

Figure 4 Relative importance of most important features (percentage values), substation scenario

both scenarios related with residential demand. This is a expected result due to the high incidence of temperature on residential energy consumption, in contrast to its low incidence on industrial energy consumption.

4.2 Experimental results on preliminary models

Performance metrics defined in Section 3.1 were used to evaluate the implementation of the one hour models as described in Section 3.2.

Tables 1–3 report the obtained results of the studied forecasting models for each scenario. The best results are reported in cells with green background.

Results reported in Table 1 for the industrial scenario indicate that three of the studied methods achieved the best results regarding the analyzed metrics. Focusing on *MAPE*, Extratreesregressor improved over MLP by 4.16%

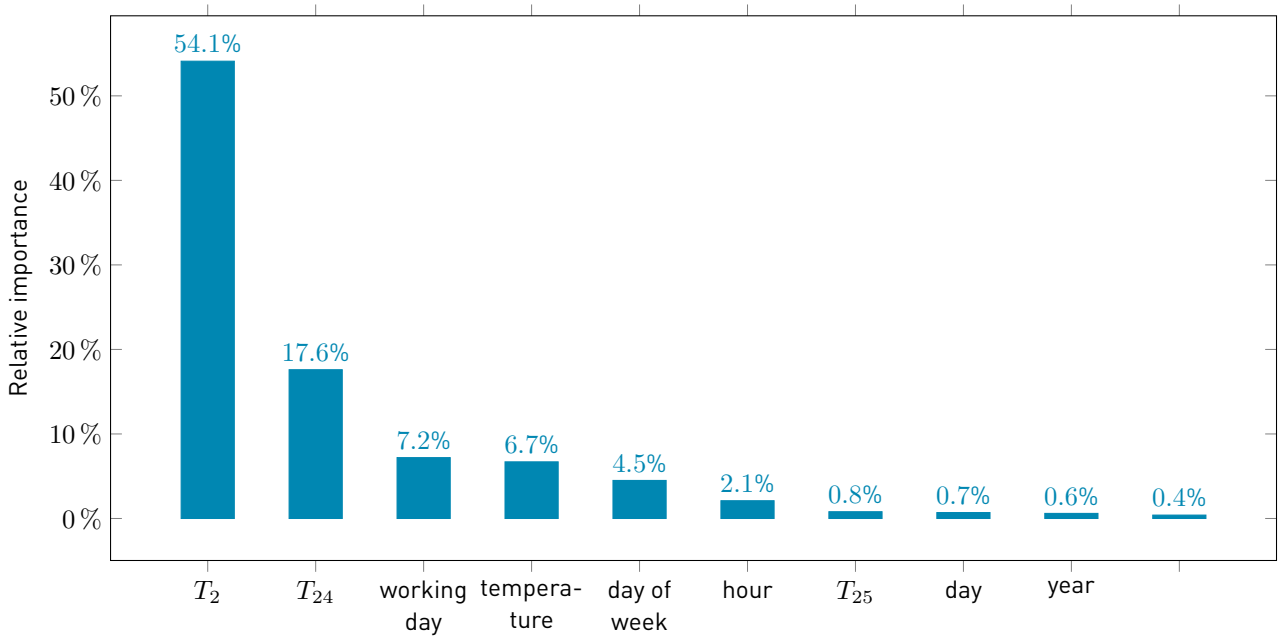

Figure 5 Relative importance of most important features (percentage values), total demand of Uruguay

and over RandomForest by 6.54%.

In turn, results reported in Table 2 for the substation scenario and in Table 3 for the total demand scenario indicate that Extratreesregressor was also the best model in both cases. Additionally, in all scenarios the training time of Extratreesregressor was approximately three times shorter than RandomForest and six times smaller than MLP.

Overall, ExtraTreesRegressor was the most effective model for forecasting the next hour, outperforming all the other methods regarding the three standard metrics studied in each scenario. According to these results, ExtraTrees was selected as the best method for showing the best performance and a low training time. Thus, in following sections M_{best} = *ExtraTreesRegressor*.

4.3 Parameter tuning

Parameter tuning techniques described in Section 3.2 were applied on the best model M_{best}.

The input for both grid search studied techniques in the three data scenarios was generated using the following values:

- *n_estimators*: [10, 50, 75, 100, 150];
- *max_features*: [auto, sqrt, log2];
- *max_depth*: [50, 100,150, 200, 250]

GridSearchCV achieved the best results with the same parameters setting for all scenarios. The best parameter setting found by the algorithm was *n_estimators*=50, *max_features*=auto and *max_depth*=250.

Regarding MAPE metric, results computed using the best configuration significantly outperformed results of the second best configuration: improvements were 14% for the industrial demand scenario, 11% in the substation demand scenario, and 12% in the total demand scenario.

4.4 Experimental results after parameter tuning

Tables 4–6 report the results of the *ExtraTreesRegressor* model before and after parameter tuning for each scenario. The best results are highlighted (cells with green background).

Results computed by the tuned configuration of *ExtraTreesRegressor* considerably improved the baseline (non-tuned) version, regarding the three studied metrics. In particular, *MAPE* reduced from 3.00% to 1.79%.

The performance improvement just demanded a negligible increase on training time increases after parameter tuning from 1.2 s to 1.7 s.

5. Experimental results of the complete model

The forecast accuracy of the final model was validated by applying $MAPE_{tot}$ a metric that extends *MAPE*. Let

Table 1 Results for each regression method in the industrial scenario

Regression method	MAE	MAPE	RMSE	Score	Time (s)
LinearRegression	127.63	3.60	176.00	0.96	1.72
Ridge	127.63	3.60	176.00	0.97	0.09
KNeighbors	180.54	5.03	253.20	0.93	0.07
RandomForest	108.20	3.21	151.54	0.98	3.10
GradientBoosting	121.97	3.38	166.17	0.97	1.99
MLP	111.08	3.13	154.23	0.97	6.21
ExtraTrees	105.44	3.00	148.61	0.99	1.21

Table 2 Results for the studied regression method in the substation scenario

Regression method	MAE	MAPE	RMSE	Score	Time (s)
LinearRegression	472.33	14.60	511.00	0.86	1.71
Ridge	473.11	14.91	521.11	0.85	0.11
KNeighbors	533.41	17.31	593.10	0.79	0.08
RandomForest	453.50	12.90	583.15	0.91	5.01
GradientBoosting	466.73	13.83	599.72	0.88	2.19
MLP	448.18	12.74	576.35	0.93	6.91
ExtraTrees	441.14	11.24	558.33	0.95	1.43

Table 3 Results for the studied regression method in the total demand scenario

Regression method	MAE	MAPE	RMSE	Score	Time (s)
LinearRegression	255.13	6.60	317.10	0.91	1.62
Ridge	262.34	6.73	296.00	0.93	0.10
KNeighbors	367.44	10.33	501.33	0.88	0.17
RandomForest	228.44	6.12	321.43	0.95	3.15
GradientBoosting	261.2	6.48	284.71	0.94	1.79
MLP	244.03	6.18	274.30	0.94	6.11
ExtraTrees	208.14	5.99	265.11	0.96	1.26

Table 4 Comparative results of ExtraTrees before and after parameter tuning

Regression method	*MAE*	*MAPE*	*RMSE*	*Score*	*Time(s)*
ExtraTrees before tuning	105.44	3.00	148.61	0.99	1.2
ExtraTrees after tuning	87.52	1.79	111.08	0.99	1.7

Table 5 Comparative results of ExtraTrees before and after parameter tuning

Regression method	MAE	MAPE	RMSE	Score	Time(s)
ExtraTrees before tuning	414.14	11.24	558.33	0.95	1.43
ExtraTrees after tuning	220.12	6.38	269.81	0.97	5.71

$MAPE_h$ be the *MAPE* value for a predicted horizon h, the extension of *MAPE* to the complete testing set is defined by Equation 10.

$$MAPE_{tot} = \frac{\sum_{i=1}^{k} MAPE_h}{k} \quad (10)$$

Tables 7–9 report the results for each one of the 24 models.

The expected behaviour is that the models trained for highly correlated hours in the future respect to the current hour, perform better than less correlated.

Table 6 Comparative results of ExtraTrees before and after parameter tuning

Regression method	MAE	MAPE	RMSE	Score	Time(s)
ExtraTrees before tuning	208.14	5.99	265.11	0.96	1.26
ExtraTrees after tuning	100.08	5.17	131.83	0.98	1.28

This fact is due to predictability, and it is enhanced when the correlation between input features and predicted values is higher.

According to Figure 1, highly correlated demand values correspond to the immediately preceding hours and from the same hours of the day before.

Analyzing the obtained results for the $MAPE_{tot}$ metric for each one of the 24 hourly models, the performance got worse from $i = 1$ to 17 and then improved from $i = 18$ to 24. These results show that highly correlated demand values performed better, as expected.

Finally, the complete model ET_{opt} was applied. A day-ahead hourly forecast demand curve was generated for each time window for the testing set and the $MAPE_{tot}$ value was calculated.

The final result for the complete model was $MAPE_{tot}$ = 2.55% in the industrial scenario, $MAPE_{tot}$ = 5.17% in the industrial scenario and $MAPE_{tot}$ = 9.09% in the substation scenario.

These results imply that the model obtained for the day ahead demand forecasting of the industrial park analyzed incurs in an error that is considered very low for most of the studies that rely on these types of models [5, 6].

For the substation scenario, there are no known previous analysis in Uruguay to compare, but considering that the group of homes connected to the substation is small, an error of $MAPE_{tot} = 9.09\%$ is considered acceptable.

For the total demand scenario, a relevant baseline for comparison is provided by the prediction models currently used by the National Administration of the Electric Market, Uruguay (ADME, adme.com.uy). According to public information reported in the ADME website, currently used prediction models have errors ($MAPE_{tot}$) between 5.00% and 7.00%, with an average of $MAPE_{tot} = 5.52\%$. The model evaluated in this article reported an error of $MAPE_{tot} = 5.17\%$, which constitutes an excellent result, improving over baseline ADME methods by 6.34%. This is a very encouraging result for total demand prediction in Uruguay.

Figures 6–8 present samples of the real demand curve and the predicted demand curve using the best model, for a subset of the testing set considered in the experiments. For the industrial demand scenario, the presented subset corresponds to the complete data from year 2017. For the substation scenario prediction, data from September 1^{st}, 2019 to December 10^{th}, 2019 are used. Finally, for the total demand scenario the subset is the complete data from year 2018.

The scenarios analyzed are representative of the studied industrial and residential demands. The results obtained with the model created were very good for all three cases.

6. Conclusions and future work

This article presented an approach to address the problem of day ahead electricity demand forecasting.

Several machine learning models were presented and studied for next hour forecasting. Recursive feature selection was applied to select the most relevant features to train the studied models. After a comparative evaluation, the best model was optimized using random search and grid search techniques.

A hybrid strategy (combining direct and recursive approaches) was built based on the optimized model for single hour prediction. It was applied to build a complete day ahead electricity demand hourly forecasting model in three scenarios: an industrial demand forecasting scenario in Spain, a residential demand forecasting scenario for a substation in Montevideo, and a total demand forecasting scenario of Uruguay, including both residential and industrial consumers.

The experimental evaluation was performed considering data from January 1^{st}, 2010 to December, 10^{th}, 2019. An extension of the *MAPE* metric was used to evaluate the complete model for the three scenarios using testing sets.

For the industrial demand scenario, the evaluation of the complete model reported a value of $MAPE_{tot} = 2.55\%$. This is a very effective prediction result, which indicates that the proposed algorithm is effective for addressing the problem of day-ahead industrial demand forecasting,

Table 7 $MAPE_{tot}$ score for each $ET_{opt,i}$ single hour model, industrial scenario

	hour											
	1	2	3	4	5	6	7	8	9	10	11	12
MAPE$_{tot}$	1.79	1.84	1.90	1.97	2.09	2.19	2.39	2.52	2.68	2.75	2.80	2.86
	hour											
	13	14	15	16	17	18	19	20	21	22	23	24
MAPE$_{tot}$	2.93	3.02	3.05	3.08	3.09	3.02	2.88	2.77	2.63	2.49	2.32	2.17

Table 8 $MAPE_{tot}$ score for each $ET_{opt,i}$ single hour model, substation scenario

	hour											
	1	2	3	4	5	6	7	8	9	10	11	12
MAPE$_{tot}$	6.38	6.55	6.77	7.02	7.45	7.80	8.5	8.98	9.55	9.80	9.98	10.19
	hour											
	13	14	15	16	17	18	19	20	21	22	23	24
MAPE$_{tot}$	10.44	10.76	10.87	10.98	11.015	10.76	10.27	9.87	9.37	8.87	8.27	7.73

Table 9 $MAPE_{tot}$ score for each $ET_{opt,i}$ single hour model, total demand scenario

	hour											
	1	2	3	4	5	6	7	8	9	10	11	12
MAPE$_{tot}$	3.63	3.73	3.85	3.99	4.24	4.44	4.85	5.11	5.43	5.58	5.68	5.80
	hour											
	13	14	15	16	17	18	19	20	21	22	23	24
MAPE$_{tot}$	5.94	6.13	6.18	6.25	6.26	6.13	5.84	5.61	5.33	5.05	4.70	4.40

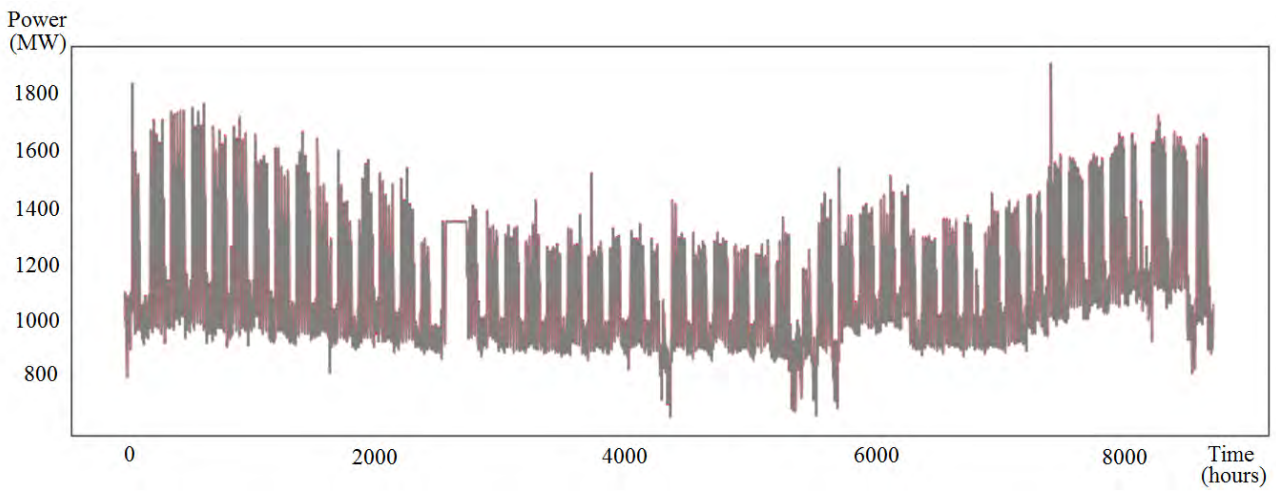

Figure 6 Predicted demand and testing data curves of industrial demand

despite of using a model that do not consider weather variables.

For the substation scenario, the evaluation of the complete model reported a value of $MAPE_{tot} = 9.09\%$. In this case, the proposed algorithm considered weather variables due to the high correlation detected between them and electricity demand. Results indicated that the complete model can predict the demand with an acceptable accuracy, in line with results from the literature, especially

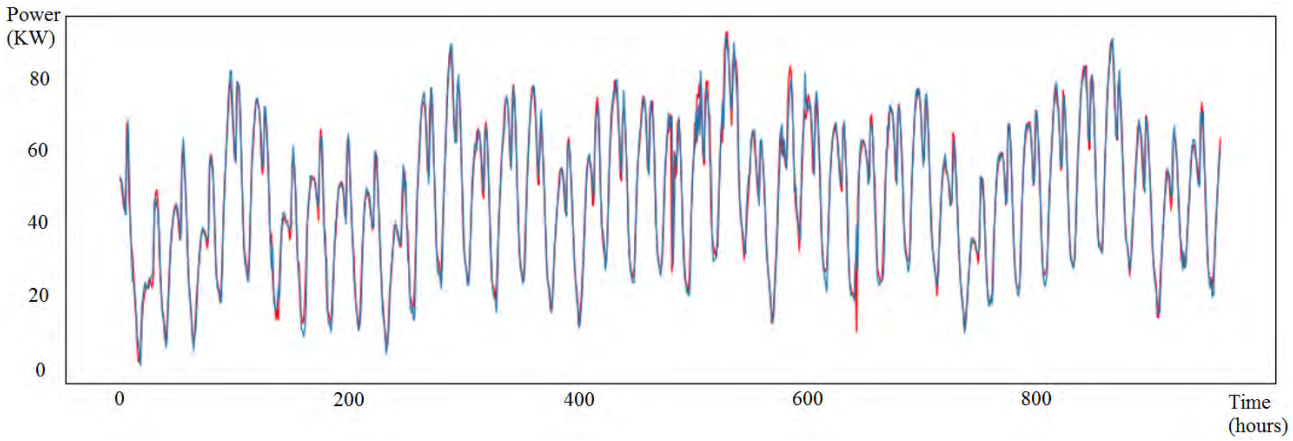

Figure 7 Predicted demand and testing data curves of substation demand

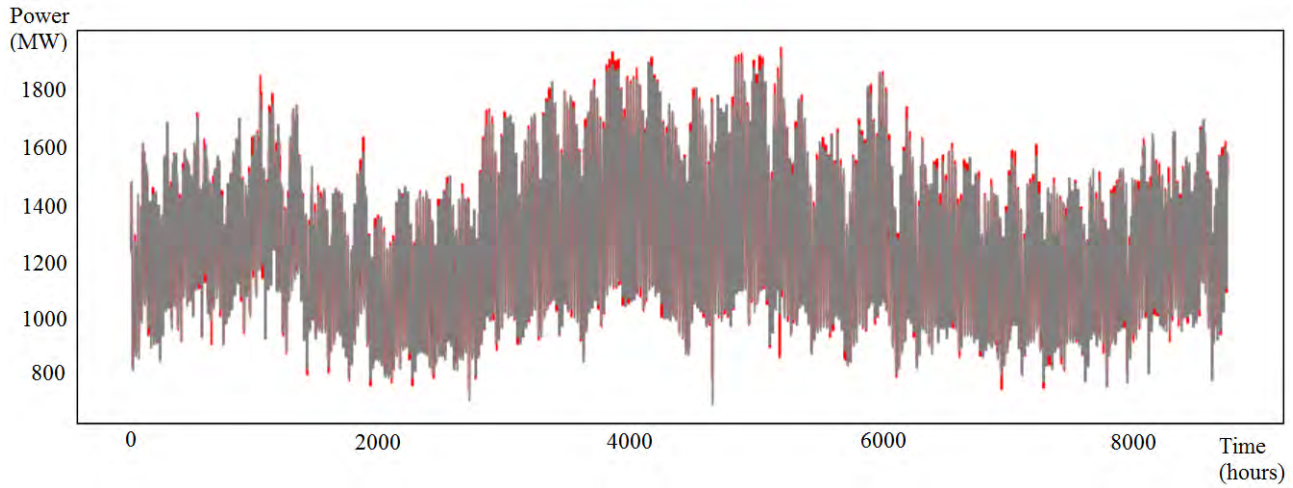

Figure 8 Predicted demand and testing data curves of total demand

considering that the variation of residential demand of a small group of houses is significantly higher than the variation of industrial demand.

Finally, the application of the complete model to the total demand scenario in Uruguay reported a value of $MAPE_{tot} = 5.17\%$. In this scenario, only one weather variable (temperature) was considered (humidity and wind speed were excluded due to low relative importance).

Results obtained are very promising considering that the models currently used in Uruguay, from the National Administration of the Electric Market has a $MAPE_{tot}$ error between 5.00% and 7.00%.

The main lines for future work are related to consider deep learning techniques (e.g., recurrent/long-short term memory neural networks) for enhancing the prediction, since they can provide accurate results in scenarios that are difficult for other simpler methods, i.e. when handling large volumes of historical data. These techniques have been successfully applied to forecasting problems with complex state structures in explanatory variables, so they can be useful tools to deal with uncertainty in electricity demand.

Another line of future work consists in enriching the studied models to generate mid-term and long-term synthetic demand scenarios that preserve the statistical structure of historical data. These kinds of models are very relevant to be included in planning and operation models based on new computational intelligence techniques such as reinforcement learning or approximate dynamic programming. Furthermore, prediction results can be applied in practice for household energy planning by using intelligent recommendation systems [37].

7. Declaration of competing interest

We declare that we have no significant competing interests including financial or non-financial, professional, or personal interests interfering with the full and objective presentation of the work described in this manuscript.

8. Acknowledgements

The research reported in this article was partly supported by CYTED-CITIES network "Ciudades Inteligentes Totalmente Integrales, Eficientes y Sostenbiles".

References

[1] A. Diniz and *et al*, "Short/mid-term hydrothermal dispatch and spot pricing for large-scale systems-the case of Brazil," in *2018 Power Systems Computation Conference (PSCC)*, Dublin, Ireland, 2018, pp. 1–7.

[2] L. Resende, M. Soares, and P. Ferreira, "Electric power load in Brazil:View on the long-term forecasting models," *Production*, vol. 28, October 8 2018. [Online]. Available: https://doi.org/10.1590/0103-6513.170081

[3] D. Lazos, A. Sproul, and M. Kay, "Optimisation of energy management in commercial buildings with weather forecasting inputs: A review," *Renewable and Sustainable Energy Reviews*, vol. 39, November 2014. [Online]. Available: https://doi.org/10.1016/j.rser.2014.07.053

[4] S. Fan, L. Chen, and W. Lee, "Machine learning based switching model for electricity load forecasting," *Energy Conversion and Management*, vol. 49, no. 6, June 2008. [Online]. Available: https://doi.org/10.1016/j.enconman.2015.07.041

[5] A. Lahouar and J. Slama, "Day-ahead load forecast using random forest and expert input selection," *Energy Conversion and Management*, vol. 103, October 2015. [Online]. Available: https://doi.org/10.1016/j.enconman.2015.07.041

[6] S. S. Ahmed, R. Thiruvengadam, S. Karrthikeyaa, and V. Vijayaraghavan, "A two-fold machine learning approach for efficient day-ahead load prediction at hourly granularity for NYC," in *FICC 2019: Advances in Information and Communication*, 2019, pp. 84–97.

[7] R. Porteiro, S. Nesmachnow, and L. Hernández, "Short term load forecasting of industrial electricity using machine learning," in *Ibero-American Congress on Information Management and Big Data (ICSC-CITIES 2019)*, 2019, pp. 146–161.

[8] J. Chavat, J. Graneri, and S. Nesmachnow, "Household energy disaggregation based on pattern consumption similarities," in *Ibero-American Congress on Information Management and Big Data (ICSC-CITIES 2019)*, 2019, pp. 54–69.

[9] R. Bellman, Ed., *Dynamic programming*, ser. Princeton Landmarks in Mathematics and Physics. United States of America: Princeton University Press, 1957.

[10] A. Soliman and A. Al-Kandari. (2010) Electrical load forecasting: modeling and model construction. [Elsevier Inc.]. [Online]. Available: https://bit.ly/2X7OdEK

[11] M. T. Hagan and S. M. Behr, "The time series approach to short term load forecasting," *IEEE Transactions on Power Systems*, vol. 2, no. 3, August 1987. [Online]. Available: https://doi.org/10.1109/TPWRS.1987.43352101

[12] J. W. Taylor and P. E. McSharry, "Short-term load forecasting methods: An evaluation based on European Data," *IEEE Transactions on Power Systems*, vol. 22, no. 4, November 2007. [Online]. Available: https://doi.org/10.1109/TPWRS.2007.907583

[13] G. Dudek, "Pattern-based local linear regression models for short-term load forecasting," *Electric Power Systems Research*, vol. 130, January 2016. [Online]. Available: https://doi.org/10.1016/j.epsr.2015.09.001

[14] C. Moreno, J. Salcedo, E. Rivas, and A. Orjuela, "A method for the monthly electricity demand forecasting in Colombia based on wavelet analysis and a nonlinear autoregressive model," *Ingeniería*, vol. 16, no. 2, July 2011. [Online]. Available: https://doi.org/10.14483/23448393.3836

[15] L. P. Caterine, K. Lin, and P. Molnár, "Electricity consumption modelling: A case of Germany," *Economic Modelling*, vol. 55, June 2016. [Online]. Available: https://doi.org/10.1016/j.econmod.2016.02.010

[16] H. Son and C. Kim, "Short-term forecasting of electricity demand for the residential sector using weather and social variables," *Resources, conservation and recycling*, vol. 123, August 2017. [Online]. Available: https://doi.org/10.1016/j.resconrec.2016.01.016

[17] C. J. Franco, J. D. Velásquez, and Y. Olaya, "Caracterización de la demanda mensual de electricidad en Colombia usando un modelo de componentes no observables," *Cuadernos de Administración*, vol. 21, no. 36, pp. 221–235, jul 2008.

[18] I. Qamber, "Peak load estimation studies in several countries," *Electric Power Systems Research*, vol. 1, no. 2, 2017.

[19] E. Burger and S. Moura, "Gated ensemble learning method for demand-side electricity load forecasting," *Energy and Buildings*, vol. 109, December 15 2015. [Online]. Available: https://doi.org/10.1016/j.enbuild.2015.10.019

[20] L. Silva, "A feature engineering approach to wind power forecasting: GEFCom 2012," *International Journal of Forecasting*, vol. 30, no. 2, April 2014. [Online]. Available: https://doi.org/10.1016/j.ijforecast.2013.07.007

[21] M. De Felice, A. Alessandri, and P. Ruti, "Electricity demand forecasting over Italy: Potential benefits using numerical weather prediction models," *Electric Power Systems Research*, vol. 104, November 2013. [Online]. Available: https://doi.org/10.1016/j.epsr.2013.06.004

[22] M. Lopez, B. Carro, and A. Sanchez, "Neural network architecture based on gradient boosting for IoT traffic prediction," *Future Generation Computer Systems*, vol. 100, November 2019. [Online]. Available: https://doi.org/10.1016/j.future.2019.05.060

[23] M. Lopez, A. Sanchez, and B. Carro, "Review of methods to predict connectivity of IoT wireless devices," *Ad Hoc & Sensor Wireless Networks*, vol. 38, no. 1-4, pp. 125–141, 2017.

[24] J. Chavat, S. Nesmachnow, and J. Graneri, "Non-intrusive energy disaggregation by detecting similarities in consumption patterns," *Revista Facultad de Ingeniería Universidad de Antioquia*, no. 98, 2020. [Online]. Available: https://doi.org/10.17533/udea.redin.20200370

[25] N. Amjady and F. Keynia, "Short-term load forecasting of power systems by combination of wavelet transform and neuro-evolutionary algorithm," *Energy*, vol. 34, no. 1, January 2009. [Online]. Available: https://doi.org/10.1016/j.energy.2008.09.020

[26] Z. Bashir and M. El-Hawary, "Applying wavelets to short-term load forecasting using pso-based neural networks," *IEEE Transactions on Power Systems*, vol. 24, no. 1, February 2009. [Online]. Available: https://doi.org/10.1016/j.energy.2008.09.020

[27] Y. Chen and *et al*, "Short-term load forecasting: Similar day-based wavelet neural networks," vol. 25, no. 1, February 2010. [Online]. Available: https://doi.org/10.1109/TPWRS.2009.2030426

[28] C. Kim, I. Yu, and Y. Song, "Kohonen neural network and wavelet transform based approach to short-term load forecasting," *Electric Power Systems Research*, vol. 63, no. 3, October 28 2002. [Online]. Available: https://doi.org/10.1016/S0378-7796(02)00097-4

[29] S. Nesmachnow, S. Bana, and R. Massobrio, "A distributed platform for big data analysis in smart cities: combining intelligent transportation systems and socioeconomic data for Montevideo, Uruguay," *EAI Endorsed Transactions on Smart Cities*, vol. 2, no. 5, December 2017. [Online]. Available: https://doi.org/10.4108/eai.19-12-2017.153478

[30] V. Jakkula and D. Cook, "Outlier detection in smart environment structured power datasets," in *2010 Sixth International Conference on*

Intelligent Environments, Kuala Lumpur, Malaysia, 2010, pp. 29–33.

[31] A. Kaleem, K. Ghori, Z. Khanzada, and N. Malik, "Address standardization using supervised machine learning," in *2011 International Conference on Computer Communication and Management*, 2011, pp. 441–445.

[32] W. McKinney, "Data structures for statistical computing in Python," in *Proceedings of the 9th Python in Science Conference (SciPy 2010)*, 2010, pp. 56–61.

[33] F. Pedregosa and *et al*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, Nov. 2011.

[34] S. Nesmachnow and S. Iturriaga, "Cluster-UY: Collaborative scientific high performance computing in Uruguay," in *International Conference on Supercomputing in Mexico (ISUM 2019)*, 2019, pp. 188–202.

[35] C. Zhang, Y. Li, Z. Yu, and F. Tian, "Feature selection of power system transient stability assessment based on random forest and recursive feature elimination," in *2016 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, Xi'an, China, 2016, pp. 1264–1268.

[36] L. Buitinck and *et al*, "API design for machine learning software: Experiences from the scikit-learn project," *ArXiv*, pp. 108–122, Sep. 2013.

[37] G. Colacurcio, S. Nesmachnow, J. Toutouh, F. Luna, and D. Rossit, "Multiobjective household energy planning using evolutionary algorithms," in *Ibero-American Congress on Information Management and Big Data (ICSC-CITIES 2019)*, 2019, pp. 269–284.