Research Article/Electrical, Electronics, and Telecommunications Engineering

http://doi.org/10.15446/ing.investig.109551

Heuristics-Based Energy Demand Forecasting with Scarce Data in the Department of Huila, Colombia

Pronóstico de la Demanda de Energía Basado en Heurística con Datos Escasos en el Departamento del Huila, Colombia

Juan J. Cuenca[1], Diego Palacios-Castro[2], and Rodolfo García[3]

ABSTRACT

Within the framework of the energy transition, electrical distribution grid operators require effective tools to predict the demand of individual users. These tools are necessary for an adequate planning of future generation resources and infrastructure modernization. However, understanding future electricity needs poses a significant challenge, especially in emerging economies, where historical data are manually collected on a monthly or bi-monthly basis and exhibit a significant amount of missing information. In response to the above, this work proposes a novel heuristics-based method for medium-term energy demand forecasting with scarce data. Qualitative and quantitative information was abstracted into a mathematical model representing the trend and noise components of historical energy consumption observations. In addition, external factors were considered as an additional layer for the mathematical model, in order to account for events that could not be foreseen by merely using the dataset. A train-test data split was proposed to iteratively search for the best parameters to predict electricity demand with respect to different categorical divisions of users (social stratum, rural or urban location, and municipality). For testing and validation, real historical data were used, as provided by the regional distribution system operator (DSO) of the department of Huila, Colomba. The results suggest a trade-off between accuracy and computational intensity, as well as the fact that a non-categorical approach leads to the algorithm with the best performance (average absolute error of 1.61%) at a low computational cost.

Keywords: demand forecasting, energy, heuristics, scarce data

RESUMEN

En el marco de la transición energética, los operadores de redes de distribución eléctrica requieren herramientas efectivas para predecir la demanda de usuarios individuales. Estas herramientas son necesarias para un planeamiento adecuado de los recursos futuros de generación y la modernización de la infraestructura. Sin embargo, entender las necesidades futuras de electricidad constituye un desafío significativo, especialmente en economías emergentes donde los datos históricos son recolectados manualmente en períodos mensuales o bimensuales y presentan una cantidad significativa de información faltante. En respuesta a esto, se propone un novedoso método basado en heurística para el pronóstico de la demanda de energía en el mediano plazo con datos escasos. Se abstrajo información cualitativa y cuantitativa en un modelo matemático que representa las componentes de tendencia y ruido en observaciones históricas de consumo de energía. Adicionalmente, se consideraron factores externos como capa adicional para el modelo matemático, en aras de dar cuenta de eventos que no podrían ser previstos solamente con el conjunto de datos. Se propuso una división de datos de entrenamiento y prueba con el fin de buscar iterativamente los mejores parámetros para predecir la demanda de electricidad respecto a diferentes divisiones categóricas de usuarios (estrato social, ubicación rural o urbana y municipio). Para realizar pruebas y validaciones, se utilizaron datos históricos reales proporcionados por el operador del sistema de distribución (OSD) regional del departamento del Huila, Colombia. Los resultados sugieren que hay una compensación entre precisión e intensidad computacional, y que un enfoque no categórico resulta en el algoritmo con un mejor desempeño (error absoluto promedio de 1.61 %) y un bajo costo computacional.

Palabras clave: pronóstico de la demanda, energía, heurística, datos escasos

Received: June 16th 2023
Accepted: September 6th 2024

Introduction

The electricity sector is undergoing significant changes due to the energy transition. The inclusion of distributed energy resources with uncertain behavior, *e.g.*, solar photovoltaic (PV), wind turbines, *etc.* (Cuenca and Hayes, 2022), the electrification of heat and transport (Mehigan *et al.*, 2022), and the development of new local energy markets and price schemes (Cuenca, Jamil, and Hayes, 2023) are changing the way we produce, transport, use, and trade electrical energy.

Within this changing paradigm, there are numerous new challenges for the operation and planning of electrical distribution networks. One of these challenges is the forecasting of energy demand. Forecasting requires the use of historical data to determine a hypothetical future

[1]Electrical engineer, Escuela Colombiana de Ingeniería, Colombia. MEng Electrical Engineering, Auckland University of Technology, New Zealand, PhD Electrical Engineering, University College Cork, Ireland. Affiliation: Researcher in Electrical and Energy Engineering at EHS Ltda., Colombia, contractor at Centro Internacional de Física (CIF), Colombia, and assistant professor of Smart Control for Energy Systems at CentraleSupélec - IETR Lab (UMR CNRS 6164), France. E-mail: juan.cuenca@ehs.com.co
[2]Affiliation: Head of the Systems and Organization Office, Electrificadora del Huila SA ESP, Colombia. E-mail: diego.palaciosc@electrohuila.co
[3]Electrical engineer, Universidad Nacional de Colombia. MSc Economic Sciences, Universidad Nacional de Colombia. PhD Engineering, Universidad Nacional de Colombia. Affiliation: Mega-projects and innovation researcher, Centro Internacional de Física (CIF), Colombia. E-mail: rgarciasi@unal.edu.co

state. This is useful for researchers, regulators, system operators, and utilities to understand how to schedule future generation resources and infrastructure upgrades (Hemmati, Hooshmand, and Taheri, 2015). In the energy sector, numerous algorithms have been developed to this effect (Klyuev *et al.*, 2022). Previous research argues that, with sufficient historical data, it is possible to approximately predict energy customer behavior.

Traditionally, historical data on energy consumption have been collected on a monthly or multi-monthly basis. The utility dispatches personnel to physically visit the customers' energy meters and record their consumption for the last billing cycle (Bimenyimana and Asemota, 2018). This is changing as we are moving towards the digital era: modern smart energy meters include measuring capabilities on smaller time steps (in the order of seconds or minutes) and utilize wired or wireless communication. This information on consumption is transmitted to the utility in close-to-real time (Bimenyimana and Asemota, 2018). Numerous nations are pioneering the rollout of smart metering, providing the necessary inputs for the functioning of forecasting algorithms in the literature.

Nonetheless, in emerging economies like Colombia, smart meter rollout is still at an early stage, and historical data on energy consumption are still collected according to tradition (*i.e.*, monthly or multi-monthly). This is especially the case with rural areas, where access difficulties may further delay the installation of smart meters. Within this frame of data scarcity, it is important to develop alternative methods for forecasting energy customer demand. In late 2021, the regional distribution system operator (DSO) of the department of Huila (Colombia), *i.e.*, Electrohuila SA ESP, opened a call for data scientists, researchers, and data enthusiasts to provide solutions to the issue of energy forecasting with scarce data. A total of 25 proposals were submitted to the Hackathon Opita Challenge call (ElectroHuila S.A. E.S.P., 2021). This manuscript reports on the most effective one of these methods.

The purpose of this study is threefold: (i) to provide a framework for data processing to abstract qualitative consumption patterns into numerical inputs; (ii) to describe the algorithm search methodology and the train-test data split in order to develop an effective heuristics-based method for forecasting electricity demand that leverages scarce real data provided by the DSO; and (iii) to report on the results of the implementation and describe potential use cases in Colombia and abroad.

Literature review

Electricity consumption forecasting has been a relevant research topic for many years. As early as 1910, experimental studies on electrical installations and the application of mathematical methods related to probability theory to calculate the future energy requirements of customers were carried out (Bunn and Farmer, 1985). Since then, and with the development of computer technology, there have been studies on the application of technocenosis, fuzzy set, game, pattern recognition, cluster analysis, and decision theories. There are numerous reviews describing the history of electricity consumption forecasting methods, which the reader is encouraged to consult (vom Scheidt *et al.*, 2020; Wei, Li, Peng, Zeng, and Lu, 2019; Ghoddusi, Creamer, and Rafizadeh, 2019; Biel and Glock, 2016).

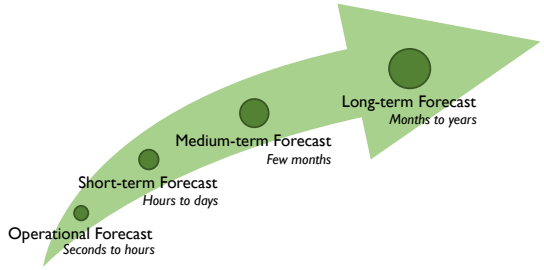

Figure 1. Classification of forecasts by lead time
Source: Authors, based on Klyuev *et al.* (2022)

Resulting from their review of the literature, (Klyuev *et al.*, 2022) recommended that, when designing forecasting methods, it is important to report not only quantitative estimates, but also qualitative features, as well as the specific conditions that make methods applicable.

Out of the numerous ways to categorize forecasting models, the most useful for the purpose of this work is by lead time (*i.e.*, the period of anticipation) (Klyuev *et al.*, 2022). As shown in Figure 1, depending on the prediction horizon, forecasts can be classified as operational, short-, medium-, and long-term (Hong and Fan, 2016). As will be discussed ahead, due to data availability, the remainder of this literature review will focus on medium-term forecasting methodologies.

Medium-term energy consumption forecasting is useful for the utility and system operators to schedule infrastructure upgrades and maintenance, plan electricity pricing, and measure the technical and economic performance of their grids (Klyuev *et al.*, 2022). Furthermore, a good knowledge of future demand is useful when defining flexibility strategies like price signaling for demand response (Honarmand, Hosseinnezhad, Hayes, Shafie-Khah, and Siano, 2021). Different approaches to the issue of energy demand forecasting are available in the literature, a selection of which is presented below.

Using deep machine learning algorithms, social and climatic factors were considered to predict the energy demand of six buildings in a city district (Yuce, Mourshed, and Rezgui, 2017). The factors were weighed to account for differential influences, and multiple regression analyses were performed. The results of this study suggest that it is possible to achieve increased forecasting accuracy in some seasons given the selected climatic factors.

The G-, Z-, and GZ-methods from statistics and time series theory were used to predict consumption by technocenosis objects while considering their individual and/or system properties (Gnatyuk, Polevoy, Kivchun, and Lutsenko, 2020). In their work, the authors introduce the autoregressive moving average (ARMA) model, the time series decomposition (TVRD) model, and the singular spectrum analysis (SSA) model.

A feature extraction algorithm was used in Meng, Niu, and Sun (2011), in which forecasting was carried out by applying

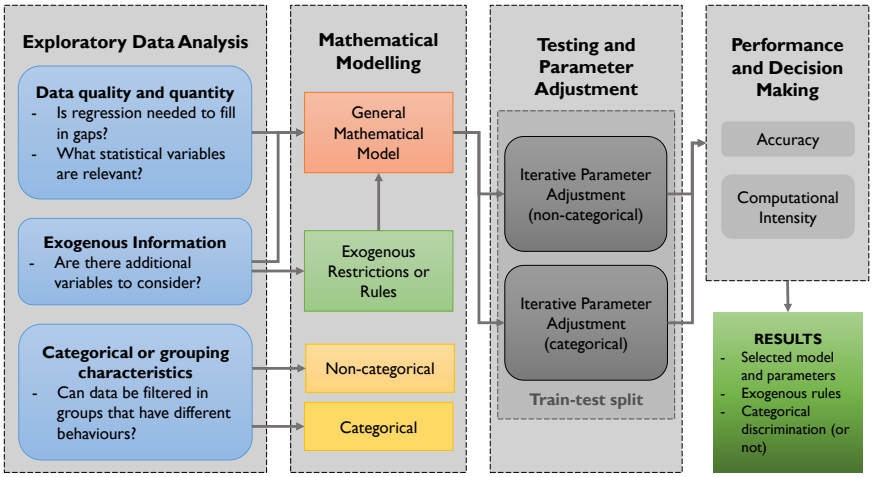

Figure 2. Structure of the proposed algorithm to select the forecasting model with scarce demand data
Source: Authors

a discrete waveform and decomposing power consumption data into a trend, a periodic component, and noise.

In Amber *et al.* (2017), a simple mathematical equation was conceived to determine future electricity usage via a multiple regression technique, considering variables such as building time, the temporal index, and surrounding temperature. These variables were were found to have an important influence on energy consumption.

Shallow neural networks were used in Shumilova, Gottman, and Starceva (2008) to address the studied problem. A three-layered perceptron with a sigmoid activation function was proposed. The input layer had 24 inputs (*i.e.*, neurons) including the power consumption of the previous month, the maximum and minimum monthly demand, the average number of hours of daylight, the monthly temperature, the number of holidays in the previous and target months, and the geographical location. The hidden layer included five neurons, and the third layer contained one neuron that gave one predicted value of consumption as the output. The authors also integrated adaptive feedback into their model to improve performance and ultimately concluded that the most accurate result is provided by a fuzzy neural network.

Classical probabilistic approaches, intelligent algorithms, and hybrid methods are used in the literature to understand trends and noise components of present demand and to ultimately forecast energy needs. As suggested in Klyuev *et al.* (2022), it is important to consider the frequency and extent of the original demand dataset to decide on the forecasting tool to be used. Monthly or multi-monthly consumption data spanning numerous years are enough to apply classical approaches, but the accuracy of intelligent algorithms depends largely on the availability of large and detailed data sources.

Considering social development, exogenous factors, the stochastic nature of human behavior, and complex macroeconomic conditions, the medium-term forecasting of energy consumption is a difficult task. The literature recommends identifying monthly electricity consumption trends using data from numerous consecutive years while considering periodic components (*e.g.*, seasonal temperature variations) and quantifying sources of noise. The above highlights the gaps in the literature on medium-term forecasting that this manuscript addresses: data scarcity, the inclusion of complex periodic components, and noise identification.

Methodology

The forecasting tool resulting from this work was obtained using a modular algorithm to select a performing model, including its structure, parameters, and rules. As seen in Figure 2, an initial data assessment is followed by the mathematical definition of the model and its variables and additional rules. Two different approaches can be applied to the mathematical model, *i.e.*, categorical or non-categorical, where the model finds or does not find different parameters for different groups of data. A train-test data split is used to analyze the performance of different parameters and structures, ultimately aiming to contrast models in terms of accuracy and computational intensity for a final decision. This section presents the details of the methodology employed in our work.

Exploratory data analysis

The first module considers a statistical analysis of the historical data available. The purpose of this initial assessment is to determine which parameters, data curation approaches, exogenous factors, and general characteristics of the dataset *should* be included in the mathematical model. First, it is important to evaluate whether the available historical data have significant missing portions that must be filled. Depending on the quantity or quality of the dataset, data curation approaches might be relevant, as discussed in Dong and Peng (2013).

Second, it is important to consider whether there are correlations between the available data and certain parameters. The following questions should be asked:

- Are there significant variations in consumption depending on the time of the year (*e.g.*, the residential electricity consumption is lower in June because it coincides with the holiday season)?

- Is temperature variation relevant (*e.g.*, the warmest month may be associated with extra electricity consumption for air conditioning)?
- Is population density relevant (*e.g.*, densely populated areas have different consumption patterns compared to rural settings)?

These questions must be framed within the socio-geo-demographic context of the case study: for different locations, social or economic contexts, different questions can be asked.

Third, we propose considering exogenous factors as part of the initial data analysis through the following questions:

- Are there any macroeconomic correlations (*e.g.*, energy consumption may be correlated to changes in energy prices)?
- Is it possible for consumption patterns to drastically change due to an exogenous event (*e.g.*, when a household is vacated, the energy consumption suddenly drops until there is a new resident)?

Lastly, provided that the dataset includes additional information (*i.e.*, not only on historical consumption), performing a comparative analysis of different categorical or grouping characteristics is very useful. If the dataset is grouped by the type of user (*e.g.*, residential, industrial, or commercial), it is possible to assess the differences in the consumption of each category. Ultimately, this illustrates the need for a differential approach that addresses the categorical nature of the dataset.

Mathematical modeling

The above-presented exploratory data analysis above sheds light on the statistical variables that may be subjected to study and are useful to build the mathematical model. Among others, values like the mean, median, variance, and standard deviation of the population or a sample can be useful to define a mathematical model.

Once identified, different mathematical relations can be drawn between variables (linear, polynomial, exponential, logarithmic, *etc.*). The selection of these relations is reserved to the model designer and involves subjective criteria, given the stochastic nature of the problem. As suggested in Lindsey (2004), mathematical model selection for stochastic problems like demand forecast is important, but secondary to the correct selection of parameters. In this vein, and for the purpose of this heuristics-based study, a linear combination of variables was selected as the mathematical model, which is shown in Equation (1). Here, a, b, and x were the selected variables, and α, β, and χ the corresponding parameters.

$$f(a,b,...,x) = (\alpha \times a) + (\beta \times b) + ... + (\chi \times x) \tag{1}$$

An intermediary block to test different mathematical relations can be added to the proposed algorithm (Figure 2). Instead of arriving at the general mathematical model directly from the exploratory data analysis, it is possible to create a loop to select an adequate formula from several candidates (*e.g.*, linear, polynomial, exponential, *etc.*) by means of a data sample. Nonetheless, given the added complexity of that approach, and since this additional block would still be influenced by the subjective criteria of the model designer, its inclusion will be addressed in future work.

After model selection, the additional rules, exceptions, or constraints resulting from the exploratory data analysis above can be superposed. It is important to correctly define and document them in terms of origin, relevance, and scope.

The resulting model can be applied to different samples of the population (*i.e.*, groups or categories). Running the selected model and adjusting parameters while considering the entire population corresponds to the non-categorical model. In contrast, the categorical model involves separately adjusting parameters for each category or group.

Testing and parameter adjustment

The above-presented mathematical model provides generality, but it is necessary to identify the parameters that will better predict the energy demand. This subsection proposes an iterative search algorithm to identify the set of parameters for a better forecast.

Train-test split approach: To evaluate different parameters for the mathematical formulation, we propose extracting a sample of the entire dataset that corresponds to the most up-to-date observations. The larger portion of the dataset (*i.e.*, the oldest observations) is used as input to train the mathematical model, and the small sample is used to test the accuracy of the predictions. This is known in the literature as a *train-test data split*, a common technique for evaluating the performance of machine learning algorithms (Medar, Rajpurohit, and Rashmi, 2017).

Error calculations: The train-test split makes it possible to evaluate a prediction using the existing dataset. To quantify the accuracy of the forecast, the percent error of the prediction $\epsilon^{\%}_{\omega}$ for the set of parameters ω is calculated. This is done through the average error between the corresponding forecast number $x^{for}_{c,\omega}$ from the training data and the mathematical model in Equation (1), and the actual observations x^{obs}_{c} from the testing sample, using Equation (2) for all customers c.

$$\epsilon^{\%}_{\omega} = \sum_{\forall c} \left(\frac{x^{for}_{c,\omega} - x^{obs}_{c}}{x^{obs}_{c}} \right) \times 100 \tag{2}$$

Sensitivity analysis: We selected the secant numerical method to find zeros in a discontinuous function. In this search algorithm, given two initial guesses ω_0 and ω_1, it is possible to iteratively obtain the best-performing value of each parameter in the set ω with a tolerance τ. The equations for the secant search algorithm are as follows:

$$\omega_j = \omega_{j-1} - \epsilon^{\%}_{\omega_{j-1}} \times \frac{\omega_{j-1} - \omega_{j-2}}{\epsilon^{\%}_{\omega_{j-1}} - \epsilon^{\%}_{\omega_{j-2}})} \tag{3}$$

$$\tau \geq \omega_j - \omega_{j-1} \tag{4}$$

Alternatively, the parameter search can be performed using an incremental approach, a bisection method, or inverse quadratic interpolation (Allen and Isaacson, 2019).

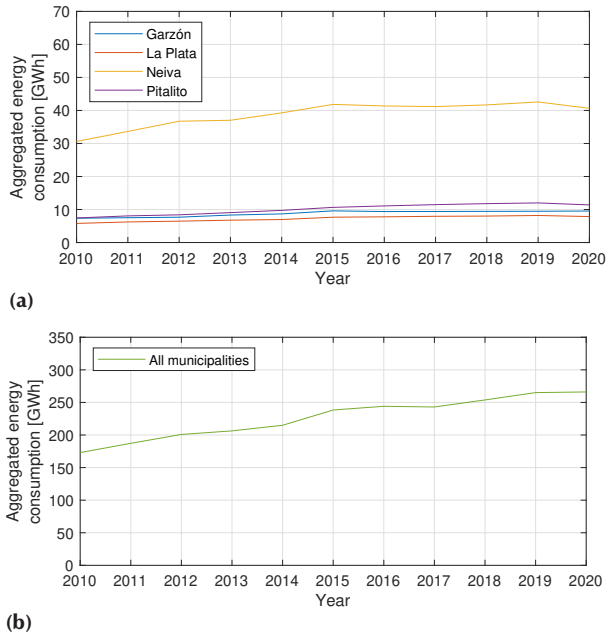

Figure 3. Evolution of the aggregated energy demand in a) the four largest cities of Huila and b) the entire department
Source: Authors

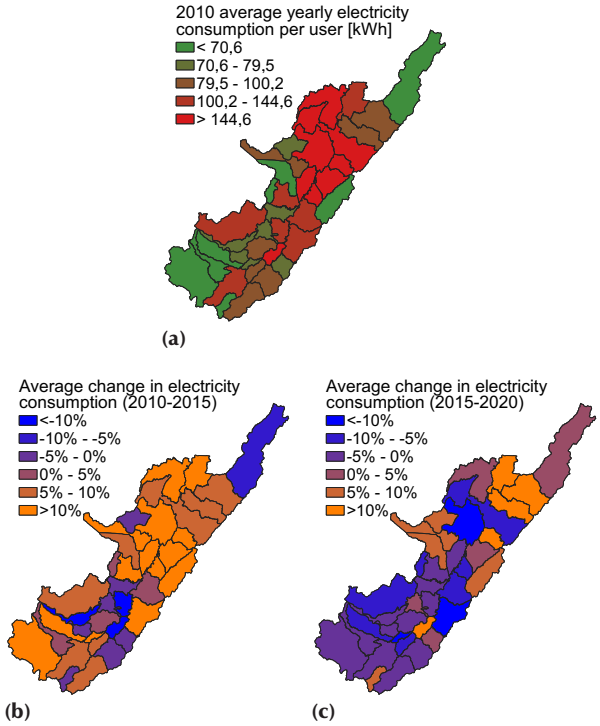

Figure 4. Yearly energy demand in the department of Huila: a) 2010 average per-customer electricity demand, b) change in electricity consumption for the 2010-2015 period, and c) change in electricity consumption for the 2015-2020 period
Source: Authors

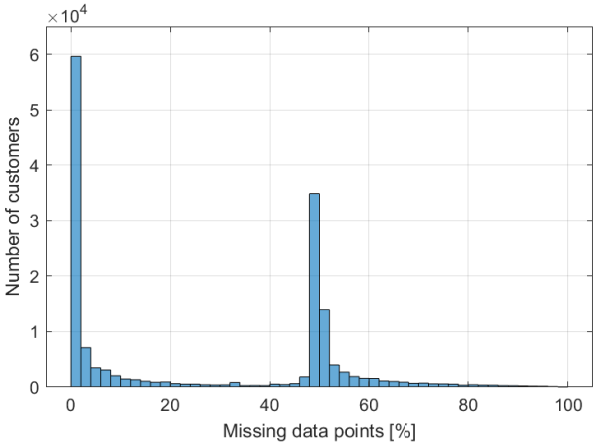

Figure 5. Histogram of missing data. Number of customers with a percentage of missing data points.
Source: Authors

Performance and decision-making

The categorical model is expected to perform better in terms of accuracy and exhibit increased computation times. This is because classification and grouping require additional computation steps, and individually analyzing every group or category also requires additional power and memory compared to the entire population. Given this trade-off between accuracy and processing times, the model designer must evaluate their context, in order to decide between a slower but more accurate categorical model or a non-categorical one with fast solutions but increased error.

Results and simulation

To test the proposed methods, the regional DSO of Huila provided a dataset containing monthly energy consumption readings from its 159 039 electricity customers in the region and other parts of the country over a 10-year period. This section presents the data analysis and simulation results. It was reported that 44 data scientists presented 25 solutions to the prediction problem during the Hackathon Opita Challenge (ElectroHuila S.A. E.S.P., 2021), and that the solution reported in this manuscript resulted in the best forecast after being tested on new data.

First, it is necessary to consider the big picture and analyze the aggregated energy demand for the department of Huila. Considering the entire population, the energy demand grows on a yearly basis (Figure 3). This growth is explained by the population increase and industrialization, and it has been extensively discussed in the literature. Based on this, a forecasting tool can simply identify the slope of the corresponding curves (*i.e.*, the growth rate) and apply it to existing observations in order to predict the demand. Nonetheless, the smooth behavior observed above is only evident when a significant amount of customers is aggregated. To extend on this, Figure 4 shows the average per-customer energy demand for 2010 as well as the 5-year growth seen in 2015 and 2020.

In Figure 4, note that the average per-customer variation in energy consumption cannot be represented by a global increase rate from Figure 3. In the same time horizon, municipalities exhibit increases or decreases amounting to the global aggregated increase. Moreover, when comparing the two time horizons, a municipality can have demand increases over a period and decreases over the next, a behavior that is not reflected in the aggregated curves in Figure 3. The global rate might be useful for a system-wide estimation of future demand, but the stochastic nature of per-customer demand requires the use of a more detailed approach.

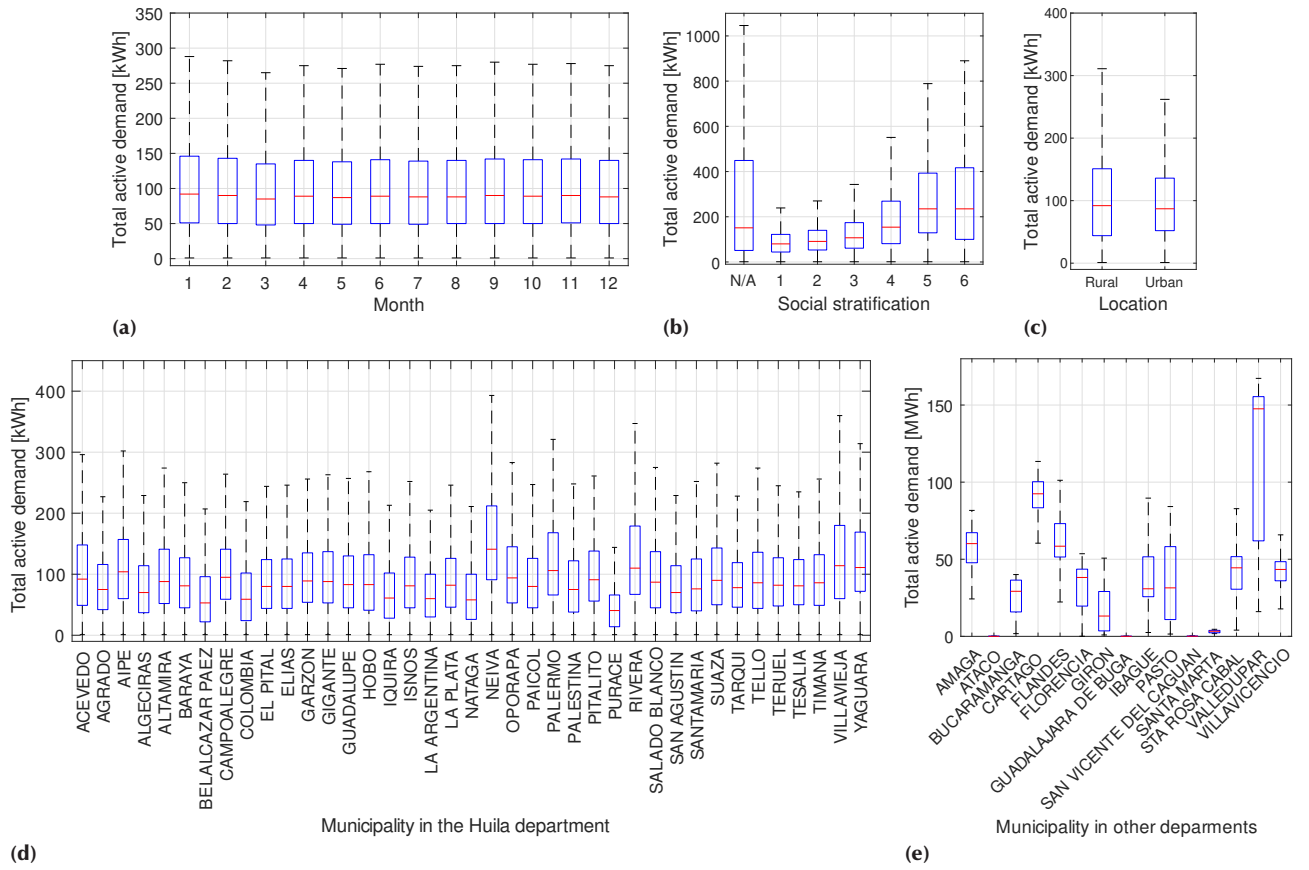

Figure 6. Categorical analysis of the population. Boxplots showing energy demand a) by month, b) by socioeconomic stratum, c) by classification (urban or rural), d) by municipality within the department, and e) by municipality outside the department. The red line represents the median, the blue box represents the 25th and 75th percentiles, and the whiskers represent the lower and upper adjacent points. Outliers are not included.
Source: Authors

Data description and curation

After organizing and filtering the information, it was necessary to curate the dataset. The main modification involved identifying and eliminating repeated data points: out of the initial 17.8 million monthly observations, approximately 29.7% were repeated values. Considering that the amount of customers over a 10-year period should amount to over 19 million observations, 36.4% of the data points were estimated to be missing. This may be due to some new customers appearing later, or some old ones disappearing before the end of the 10-year horizon. Figure 5 presents a histogram with the percentage of missing data points as a proportion between the dates of the first and last observations.

There are two peaks in Figure 5 at 0-5 and 50-55%. This suggests that (i) most customers have at least some data points missing, and (ii) more than 35 000 customers have bimonthly observations, in contrast with the initially reported monthly dataset.

Furthermore, a significant number of customers with monthly observations have between 5 and 20% of missing data points, and those with bi-monthly observations have 55-80%. This highlights the importance of an approach that can deal with data scarcity.

Once the dataset had been homogenized, a categorical analysis was performed. The DSO differentiates customers by social stratum, location (*i.e.*, rural or urban), and municipality. The following questions must be considered:

- Is income disparity relevant?
- Is population density relevant?
- Can data be filtered in groups that exhibit different behaviors?

To evaluate this, Figure 6 shows the boxplots of all the data per category, as well as a monthly consumption analysis.

A relatively small difference between the energy consumption for different times of the year is observed. The months with the largest and smallest average consumptions are January and March, reporting medians of 92 and 85 kWh, respectively. Regarding the proportions of the quartiles in Figure 6a, it is important to note that there are significant variation values, as represented by the whiskers of the boxplot. That is to say that, when considering the entire population, the data variation is significant.

As for the different categories in Figure 6, it is clear that a customer belonging to one or the other exhibits a significantly different behavior. This suggests the value of employing a segregated approach that considers the categorization of the population. The most striking categorical difference is observed in social stratification (Figure 6b), where the median, the 25th and 75th percentiles, and the upper adjacent points of customers categorized as low-income (*i.e.*, social strata 1 and 2) are three times lower than those of high-income customers (social strata 5 and 6). Users that are not categorized into a social stratum are understood to be commercial, industrial,

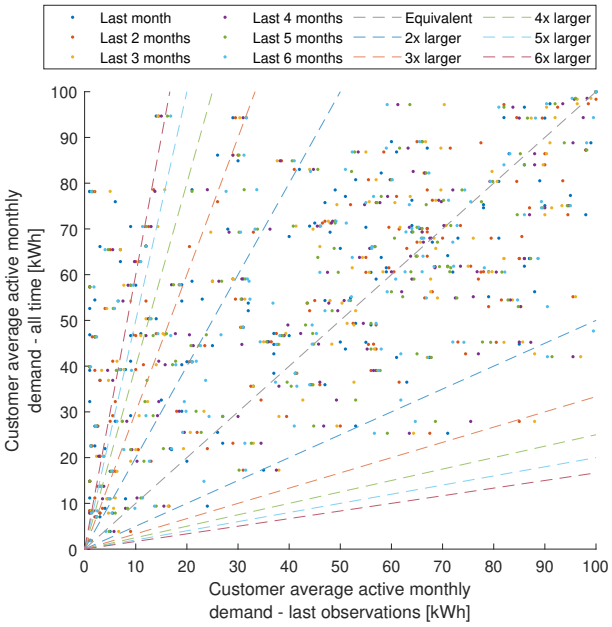

Figure 7. Scatter plot to determine recent changes in consumption patterns. The *y*-axis presents the average consumption of a customer based on all observations, and the *x*-axis corresponds to the last few months.
Source: Authors

or agricultural in nature, which explains the high variability in their energy consumption.

Different municipalities exhibit distinct energy consumption patterns. Neiva, the capital of Huila, reports the highest median energy consumption and has the largest observation variability. In addition, note the difference in scales between Figures 6d and 6e. Customers located outside Huila exhibit consumption patterns many times larger than those within the department. While this is not clarified in the documentation of the original dataset, it may be due to the existence of large-scale customers outside the department which are not physically connected to the grids of the DSO but employ it as a service provider.

The last consideration of the data analysis was whether there were exogenous factors to consider in the mathematical models. The assumption made in this work was that consumption patterns may change in the last observations. Figure 7 shows a scatter plot of the average values for the last one to six observations and the average value of all observations, given a random sample of 100 customers.

In this Figure, the diagonal grey-dashed line represents an equivalent value between the last observations and the all-time averages, which means that values close to the diagonal correspond to customers whose consumption patterns have not changed in the last few months. In contrast, the other color-dashed lines (*i.e.*, navy, orange, green, blue, and burgundy) represent how many times larger or smaller the short-term average is relative to the all-time value. Data points located between the blue-dashed line and the orange-dashed one counting counter-clockwise from the diagonal show short-term observations that are two to three times smaller than the all-time average, suggesting a change in consumption patterns.

Note that a significant number of points is located counter-clockwise from the burgundy-dashed line, representing

Algorithm 1 Parameter selection pseudo-code

Get C ▷ Number of categories
Get ω_0 and ω_1 ▷ Two initial parameter guesses
Get τ ▷ Error tolerance
Get i_{max} ▷ Maximum iterations allowed
$P_{train} = P \cap P_{test}$ ▷ Split population for training-testing
Calculate forecast for initial guesses (Equation (1)):
$x^{for}_{\omega_0}, \leftarrow f(P_{train}, \omega_0)$
$x^{for}_{\omega_1} \leftarrow f(P_{train}, \omega_1)$
$x^{obs} \leftarrow P_{test}$ ▷ Get observations from testing population
Calculate error for initial guesses (Equation (2)):
$\epsilon_0 \leftarrow g(x^{for}_{\omega_0}, x^{obs})$
$\epsilon_1 \leftarrow g(x^{for}_{\omega_1}, x^{obs})$
for $c \leftarrow 1$ **to** C **do** ▷ Do this for each category
  for $i \leftarrow 1$ **to** i_{max} **do** ▷ Do this for each iteration
  Calculate new guess for next iteration (Equation (3)):
    $\omega_i \leftarrow h(\omega_{i-1}, \omega_{i-2}, \epsilon_{i-1}, \epsilon_{i-2})$
    $T \leftarrow \omega_i - \omega_{i-1}$ ▷ Calculate tolerance (Equation (4))
    if $T \leq \tau$ **then**
      $\omega_c \leftarrow \omega_i$ ▷ Store parameters for category c
      break ▷ Stop the for loop, parameters found
    else
      $x^{for}_{\omega_i}, \leftarrow f(P_{train}, \omega_i)$ ▷ Equation (1) for ω_i
      $\epsilon_i \leftarrow g(x^{for}_{\omega_i}, x^{obs})$ ▷ Equation (2) for ω_i
    end if
  end for
end for

short-term observations more than six times smaller than the all-time average. This procedure was followed for five different 100-customer random samples, with equivalent results. While there is no additional information, this evinces that changes in consumption patterns are an important factor for energy forecasting.

The lower portion of Figure 7 (*i.e.*, counting clockwise from the grey-dashed line) includes significantly fewer relative observations. This means that, while it is common for the average consumption to be significantly reduced in the last few months compared to all observations, the opposite is not often the case. Consumption patterns change towards a net decrease, which may be due to short periods of residential vacancy or industry stall periods. Considering that these low-consumption periods are expected to be short in the scale of the dataset (a few months of the 10-year period), any increase in consumption caused by re-occupancy is diluted by a longer average occupancy over previous months.

Based on the exploratory analysis of the information provided by the DSO, we decided to include the monthly median and standard deviation of the user in the mathematical model. These variables represent the trends and noise, respectively. The categorical and non-categorical models were tested to assess the trade-off between accuracy and computational intensity. Each parameter associated with these variables is presented in the next subsection. In addition, an exogenous rule to account for changes in consumption patterns was included: if the average energy consumption of a customer for the last six observations was more than six times larger/smaller than the average of all their observations, a change in consumption pattern

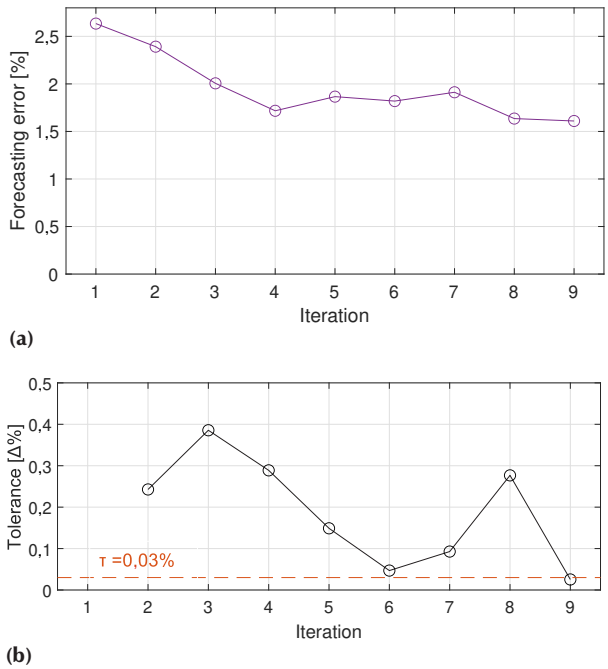

Figure 8. Results for the non-categorical parameter selection algorithm: a) forecast error for the train-test split and b) its corresponding tolerance
Source: Authors

was assumed { these values represent the most critical change observed in Figure 7. In this case, the forecast from the model was replaced by the average of these last six observations.

Parameter adjustment

The parameters for each categorical model (and its relevant categories) were calculated using an iterative search algorithm based on the secant method, whose pseudo-code is presented in Algorithm 1. This algorithm was executed for every categorical division outlined in the previous sections.

Depending on how the population is split, the runtime of Algorithm 1 can be significant. We recommend carefully selecting the size of the testing population. This is especially important for categorical models, as the algorithm includes a for loop to calculate the parameters of each category. Computational intensity is therefore linked to the testing population size and the number of categories.

The least computationally intensive scenario corresponded to the non-categorical model, whose parameters were calculated for the entire population. To illustrate this, Figure 8 presents the error and tolerance of each iteration in obtaining the non-categorical parameters. The testing portion of the population was January, February, and March 2020, and the training portion included every previous observation (*i.e.*, from January 2010 until December 2019).

The behavior in Figure 8 suggests that the search algorithm correctly identifies a solution with a low forecasting error given the selected tolerance. However, note that iteration 6 almost fulfilled the condition $\tau = 0.03\%$. The discontinuous nature of the error function from Equations (1), (2), and (3) leads to locally minimum tolerances and, hence, to premature solutions. Future developments of this work should include additional iterations to confirm whether there is a nearby local minimum with a better performance.

Algorithm selection

The algorithm was coded in the SciLab 6.1.1 open-access software and run using a desktop with 16 GB RAM and an Intel Core i5-8400 CPU. The accuracy and computational intensity resulting from the train-test split analysis are shown in Table 1.

Table 1. Comparison of the results obtained from the train-test split for different categories

Categorical Division	Num. Categ.	Total Iter.	Run time [h] Train	Test	Train-test Error [%]
None	1	9	14.5	1.73	1.61
Social stratum	7	39	61.4	5.72	1.59
Location	2	15	23.13	2.56	1.64
In/out of Huila	2	23	36.88	3.14	1.58
Municipality	54	342	457.2*	37.6	1.43

* This long parameter search was run for two categories at a time.

Source: Authors

The previously hypothesized trade-off between computational intensity and accuracy was confirmed by the results in Table 1. The best-performing algorithm involves segregating parameters for each municipality, but this comes at a very high computational cost: 19 days of training time. Assigning equal value to accuracy and computational intensity, it is possible to compare the relative computational cost per error unit. In the design constraints, the DSO specified that computational intensity was an important decision variable, so the non-categorical model was selected, with an average error of 1.61% in the train-test simulation and a total training time of 14.5 hours.

The methods presented in this work provide the DSO with the flexibility to decide which categorical division to select according to its particular needs regarding accuracy and computational intensity. This is reserved as a decision variable for future applications of this work.

Discussion

The review by Klyuev *et al.* (2022) shows that the absolute percentage error of medium-term electricity demand predictions oscillates between 7.95 and 15.92%. For the sake of comparison, the absolute percent errors reported by other authors are compared against our results in Table 2. The authors referenced therein did not report execution times, which are difficult to benchmark in any case because each study uses different population sizes, has varying degrees of missing data, and considers more or less historical data.

This tool can be extended for application by other regional and national DSOs. Considering the computational constraints, it is possible to obtain a more accurate solution via the train-test split of historical data. This is also determined by the purpose of the forecast (*e.g.*, if it aims to schedule and prioritize infrastructure upgrades, accuracy is more important than computational intensity (Migliavacca *et al.*, 2021)).

The work of Schafer (1999) suggests that missing data amounting to 5% or less can be tolerated in statistical

Table 2. Benchmarking against other methods in the literature

Reference	Method	Dataset (granularity)	Reported Error [%]
Proposed	Heuristics	10 years (1-2 month)	1.61
(Yuce *et al.*, 2017)	Deep ML	1.5 years (not reported)	4.5-8.8
(Meng *et al.*, 2011)	Feature extraction	2 years (1 month)	2.7-2.8
(Amber *et al.*, 2017)	Multiple regression	5 years (not reported)	8.6-9.8

* ML = machine learning.

Source: Authors

analysis. Furthermore, Bennett (2001) argues that more than 10% of missing data points causes any statistical analysis to be biased. Thereupon, as future work, we propose the use of regression to fill large gaps in energy consumption datasets (*e.g.*, there are more than 36% missing data points in the set used in this work). Numerous approaches for data regression are suggested in Dong and Peng (2013) in relation to the missing data mechanism or its origin.

The historical consumption dataset used in this article includes one temporal (month), two spatial (municipality and urban/rural location), and one socioeconomic feature (social stratum) for individual consumers. Note that the social stratum is an official classification implemented by the Colombian government to group dwellings with similar social and economic characteristics within a delimited area, and it does not include useful information about *individual* household characteristics (Chica-Olmo, Sánchez, and Sepúlveda-Murillo, 2020). In this regard, we propose treating social and economic factors independently and on a larger spectrum, *i.e.*, economic factors such as family employment, household income, debt capacity, and savings; and social/demographic factors like family size, behavior towards the environment, and population density, among others. These can be used to build and select models that better represent individual behavior (Yuce *et al.*, 2017). Due to lack of data, this will be studied in future work.

Future applications of this work could include the definition of dynamic tariffs that account not only for generation resource availability but also for forecast demand scenarios (Ochoa, Dent, and Harrison, 2010). Flexibility resources, demand response, and infrastructure upgrades could be optimally planned if the future demand is known (Capitanescu, Ochoa, Margossian, and Hatziargyriou, 2015). The hosting capacity of distribution networks largely depends on operational states that rely on an accurate representation of future demand (Ochoa *et al.*, 2010). Ultimately, a good demand prediction is key to assess the reliability and resilience of modern distribution grids (Escalera, Hayes, and Prodanović, 2018).

Conclusions

This manuscript presents a historical demand data processing framework for medium-term electricity consumption forecasting. Qualitative and quantitative patterns were abstracted to build a mathematical model, which was later tested given different categorical divisions. The best-performing method was selected while considering the trade-off between accuracy and computational intensity. This method was tested using real data from the regional DSO of the department of Huila, Colombia. This DSO reported that applying the selected algorithm resulted in a forecast that was at least 3% more accurate than other solutions regarding the real customer demand (*i.e.*, data not used to train or test the algorithm).

Through heuristics, simple statistical quantities of the population and samples of it can be used to build a robust model with accurate outputs. It was found that it is necessary to account for exogenous factors. In this case, changes in consumption patterns played an important role in forecasting energy demand.

Opportunities for future work include filling the gaps in the dataset, especially considering that a significant amount of customers exhibit a bimonthly cycle of observations. Moreover, additional categories could be created for data filtering. By mixing two categories, subcategories that better inform customer behavior could be obtained (*e.g.*, by mixing social stratification and the type of location, consumption patterns can be better represented). However, this would be limited by computation time constraints, as it represents the inclusion of subcategories, which would considerably increase the size of the problem.

Acknowledgments

The authors would like to give special thanks to Electrohuila SA ESP for providing the data used in this work for testing and validation, and for funding and organizing the Hackathon Opita Challenge of 2021, from which the work presented herein is a major output. Special thanks to M. Rouillé-Tamayo for her support, as well as to Centro Internacional de Física and EHS Ltda. in Colombia for their financial and infrastructural support in writing this manuscript.

Author contributions

Cuenca J. conceived the idea and did the background research. *Palacios-Castro D.* collected the data. *García R.* provided critical feedback. *Cuenca J.* led the writing process and wrote the main part of the manuscript, to which all authors contributed.

Conflicts of interest

There are no conflicts of interest to declare.

References

Allen, M., and Isaacson, E. (2019). *Numerical analysis for applied science*. Wiley. https://books.google.be/books?id=PpB9cj0xQAQC.

Amber, K. P., Aslam, M. W., Mahmood, A., Kousar, A., Younis, M. Y., Akbar, B., ... Hussain, S. H. (2017). Energy consumption forecasting for university sector buildings. *Enegries*, *10*(10). https://doi.org/10.3390/en10101579

Bennett, D. A. (2001). How can i deal with missing data in my study? *Australian and New Zealand Journal of Public Health*, *25*(5), 464-469. https://doi.org/10.1111/j.1467-842X.2001.tb00294.x

Biel, K., and Glock, C. H. (2016). Systematic literature review of decision support models for energy-efficient production planning. *Computers & Industrial Engineering*, *101*, 243-259. https://doi.org/10.1016/j.cie.2016.08.021

Bimenyimana, S., and Asemota, G. N. O. (2018). Traditional vs smart electricity metering systems: A brief overview. *Journal of Marketing and Consumer Research*, *46*, 1-7. https://www.iiste.org/Journals/index.php/JMCR/article/view/42505/43773. (Accessed: 2023-03-02)

Bunn, D., and Farmer, E. (1985). *Comparative models for electrical load forecasting*. Wiley. https://www.osti.gov/biblio/6256333. (Accessed: 2023-03-02)

Capitanescu, F., Ochoa, L. F., Margossian, H., and Hatziargyriou, N. D. (2015). Assessing the potential of network reconfiguration to improve distributed generation hosting capacity in active distribution systems. *IEEE Transactions on Power Systems*, *30*(1), 346-356. https://doi.org/10.1109/TPWRS.2014.2320895

Chica-Olmo, J., Sánchez, A., and Sepúlveda-Murillo, F. H. (2020). Assessing colombia's policy of socio-economic stratification: An intra-city study of self-reported quality of life. *Cities*, *97*, 102560. https://www.sciencedirect.com/science/article/pii/S0264275119312995. https://doi.org/10.1016/j.cities.2019.102560

Cuenca, J. J., and Hayes, B. P. (2022). Non-bias allocation of export capacity for distribution network planning with high distributed energy resource integration. *IEEE Transactions on Power Systems*, *37*(4), 3026-3035. https://doi.org/10.1109/TPWRS.2021.3124999

Cuenca, J. J., Jamil, E., and Hayes, B. P. (2023). Revenue-based allocation of electricity network charges for future distribution networks. *IEEE Transactions on Power Systems*, *38*(2), 1728-1738. https://doi.org/10.1109/TPWRS.2022.3176186

Dong, Y., and Peng, C.-Y. J. (2013, May 14). Principled missing data methods for researchers. *SpringerPlus*, *2*(1), 222. https://doi.org/10.1186/2193-1801-2-222

ElectroHuila S.A. E.S.P. (2021). *Hackatón OpitaChallenge*. https://reto.electrohuila.com.co/. (Accessed: 2023-03-02)

Escalera, A., Hayes, B., and Prodanović, M. (2018). A survey of reliability assessment techniques for modern distribution networks. *Renewable and Sustainable Energy Reviews*, *91*, 344-357. https://doi.org/10.1016/j.rser.2018.02.031

Ghoddusi, H., Creamer, G. G., and Rafizadeh, N. (2019). Machine learning in energy economics and finance: A review. *Energy Economics*, *81*, 709-727. https://doi.org/10.1016/j.eneco.2019.05.006

Gnatyuk, V. I., Polevoy, S. A., Kivchun, O. R., and Lutsenko, D. V. (2020, apr). Applying the potentiating procedure for optimal management of power consumption of technocenose. *IOP Conference Series: Materials Science and Engineering*, *837*(1), 012001. https://doi.org/10.1088/1757-899X/837/1/012001

Hemmati, R., Hooshmand, R.-A., and Taheri, N. (2015). Distribution network expansion planning and dg placement in the presence of uncertainties. *International Journal of Electrical Power & Energy Systems*, *73*, 665-673. https://doi.org/10.1016/j.ijepes.2015.05.024

Honarmand, M. E., Hosseinnezhad, V., Hayes, B., Shafie-Khah, M., and Siano, P. (2021). An overview of demand response: From its origins to the smart energy community. *IEEE Access*, *9*, 96851-96876. https://doi.org/10.1109/ACCESS.2021.3094090

Hong, T., and Fan, S. (2016). Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting*, *32*(3), 914-938. https://doi.org/10.1016/j.ijforecast.2015.11.011

Klyuev, R. V., Morgoev, I. D., Morgoeva, A. D., Gavrina, O. A., Martyushev, N. V., Efremenkov, E. A., and Mengxu, Q. (2022). Methods of forecasting electric energy consumption: A literature review. *Energies*, *15*(23). https://doi.org/10.3390/en15238919

Lindsey, J. K. (2004). *Statistical analysis of stochastic processes in time*. Cambridge University Press. https://doi.org/10.1017/CBO9780511617164

Medar, R., Rajpurohit, V. S., and Rashmi, B. (2017). Impact of training and testing data splits on accuracy of time series forecasting in machine learning [conference paper]. In *2017 international conference on computing, communication, control and automation (iccubea)*. https://doi.org/10.1109/ICCUBEA.2017.8463779

Mehigan, L., Zehir, M. A., Cuenca, J. J., Sengor, I., Geaney, C., and Hayes, B. P. (2022). Synergies between low carbon technologies in a large-scale mv/lv distribution system. *IEEE Access*, *10*, 88655-88666. https://doi.org/10.1109/ACCESS.2022.3199872

Meng, M., Niu, D., and Sun, W. (2011). Forecasting monthly electric energy consumption using feature extraction. *Energies*, *4*(10), 1495–1507. https://doi.org/10.3390/en4101495

Migliavacca, G., Rossi, M., Siface, D., Marzoli, M., Ergun, H., Rodríguez-Sánchez, R., ... Morch, A. (2021). The innovative flexplan grid-planning methodology: How storage and flexible resources could help in debottlenecking the european system. *Energies*, *14*(4). https://doi.org/10.3390/en14041194

Ochoa, L. F., Dent, C. J., and Harrison, G. P. (2010). Distribution network capacity assessment: Variable dg and active networks. *IEEE Transactions on Power Systems*, *25*(1), 87-95. https://doi.org/10.1109/TPWRS.2009.2031223

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, *8*(1), 3–15. https://doi.org/10.1191/096228099671525676

Shumilova, G., Gottman, N., and Starceva, T. (2008). Forecasting of electrical loads in the operational management of electric power systems based on neural network structures. *KNC UrO RAS: Syktyvkar, Russia*, 85.

vom Scheidt, F., Medinová, H., Ludwig, N., Richter, B., Staudt, P., and Weinhardt, C. (2020). Data analytics in the electricity sector. a quantitative and qualitative literature review. *Energy and AI*, *1*, 100009. https://doi.org/10.1016/j.egyai.2020.100009

Wei, N., Li, C., Peng, X., Zeng, F., and Lu, X. (2019). Conventional models and artificial intelligence-based models for energy consumption forecasting: A review. *Journal of Petroleum Science and Engineering*, *181*, 106187. https://doi.org/10.1016/j.petrol.2019.106187

Yuce, B., Mourshed, M., and Rezgui, Y. (2017). A smart forecasting approach to district energy management. *Energies*, *10*(8). https://doi.org/10.3390/en10081073