Data Article

Distribution level electric current consumption and meteorological data set of the east region of Paraguay

Gustavo Velázquez\textsuperscript{a}, Félix Morales\textsuperscript{a}, Miguel García-Torres\textsuperscript{a,b,*}, Francisco Gómez-Vela\textsuperscript{b}, Federico Divina\textsuperscript{b}, José Luis Vázquez Noguera\textsuperscript{a}, Federico Daumas-Ladouce\textsuperscript{a}, Carlos Sauer Ayala\textsuperscript{c}, Diego P. Pinto-Roa\textsuperscript{d}, Pedro E. Gardel-Sotomayor\textsuperscript{a,e}, Julio César Mello Román\textsuperscript{a,f}

\textsuperscript{a} Computer Engineer Department, Universidad Americana, Asunción 1029, Paraguay
\textsuperscript{b} Data Science and Big Data Lab, Universidad Pablo de Olavide, Seville ES-41013, Spain
\textsuperscript{c} Facultad de Ingeniería, Universidad Nacional de Asunción, San Lorenzo 111421, Paraguay
\textsuperscript{d} Facultad Politécnica, Universidad Nacional de Asunción, San Lorenzo 111421, Paraguay
\textsuperscript{e} Facultad de Ciencias y Tecnología, Universidad Católica Nuestra Señora de la Asunción, Campus Alto Páraná, Paraguay
\textsuperscript{f} Facultad de Ciencias y Tecnología, Universidad Católica Nuestra Señora de la Asunción, Campus Alto Paraná, Paraguay

\section*{ARTICLE INFO}

Article history:
Received 13 July 2021
Revised 16 November 2021
Accepted 8 December 2021
Available online 10 December 2021

Keywords:
Electric current consumption
Meteorological data
Forecasting
Machine learning

\section*{ABSTRACT}

This paper presents a data set with information on meteorological data and electricity consumption in the department of Alto Paraná, Paraguay. The meteorological data were registered every three hours at the Aeropuerto Guarani, Department of Alto Paraná, which belongs to the Dirección Nacional de Aeronáutica Civil of Paraguay. The final data consists of a total of 22,445 records of temperature, relative humidity, wind speed and atmospheric pressure. On the other hand, the electrical energy consumption data set contains a total of 1,848,947 records, all of them coming from the one hundred and fifteen feeders located throughout the Alto Paraná region of Paraguay. Electrical energy consumption data was provided by Administración Nacional de Electricidad (ANDE).

* Corresponding author at: Data Science and Big Data Lab, Universidad Pablo de Olavide, ES-41013 Seville, Spain.
E-mail address: mgarciat@upo.es (M. García-Torres).

https://doi.org/10.1016/j.dib.2021.107699
2352-3409/© 2021 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)
The analysis of this data can yield insights regarding the energy consumption in the area.

© 2021 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

### Specifications Table

| Subject               | Energy |
|-----------------------|--------|
| Specific subject area | Electricity consumption and meteorological data. |
| Type of data          | Table. |
| How data were acquired| Electricity consumption and meteorological data were extracted from spreadsheets provided by the Administración Nacional de Electricidad (ANDE) and the Dirección Nacional de Aeronáutica Civil del Paraguay (DINAC) respectively. |
| Data format           | Raw, Analyzed and Processed in comma-separated values (CSV) format. |
| Parameters for data collection | All feeders in the electricity consumption set contains at least 90% of data, those with less data were automatically discarded. The data was anonymized by replacing the real name of the feeders. |
| Description of data collection | The data set provides information of electricity consumption and meteorological data of the region of Alto Paraná, Paraguay. Electricity consumption set includes datetime, substation, feeder and consumption with an hourly frequency. Meteorological data includes datetime, temperature, relative humidity, wind speed and atmospheric pressure at the station level with a frequency of every three hours. |
| Data source location  | Alto Paraná, Paraguay (25.6076° S, 54.9612° W). |
| Data accessibility    | Repository name: Electric current consumption and meteorological data of Alto Paraná, Paraguay. |
|                       | Data identification number: 10.17632/hzfwzzsk8f4 |
|                       | Direct URL to data: https://data.mendeley.com/datasets/hzfwzzsk8f/4 |

### Value of the Data

- The data provided makes it possible the investigation of the electricity balance, as well as the analysis of customer behaviour at a distribution feeder level.
- The analysis of electric current consumption allows to design and develop energy saving strategies.
- The analysis of electric data can be useful to distribution network operation and expansion planning.
- The impact of electrical tariff variations and demand side management strategies can be analyzed with the aid of this dataset.
- The data can be modeled with predictive and descriptive models.
- The data can be used for benchmarking to assess the quality of predictive and descriptive algorithms.

### 1. Data Description

The data set provides information of electricity consumption and meteorological data of the region of Alto Paraná, Paraguay. It is presented in 5 files:

1. electricity-consumption-raw.csv
2. electricity-consumption-processed.csv
3. substations-geographical-location.csv
4. meteorological-raw.csv
5. meteorological-processed.csv

The data that each file contains is presented in the next sections.

Electricity consumption sets includes date, time, substation, feeder and consumption with an hourly frequency. Meteorological data sets includes date, time, temperature, relative humidity, wind speed and atmospheric pressure at the station level with a three hours frequency. This data set can be used to train and validate the performance of machine learning algorithms used in regression and classification as well as modelling, simulation and optimization problems related to energy consumption and climate.

1.1. Geographical and climate overview

Fig. 1 shows a visual representation of the geographical location of Alto Paraná, which is located in the east of the Eastern Region of Paraguay. Grassi et al [5] reported that, according to the general circulation of the atmosphere, the region is located in the domain of the western edge of the subtropical anticyclone of the South Atlantic; thus, it is bathed by warm and humid conditions.
winds from the Atlantic Ocean, for most of the year, through the prevailing northeast winds. The department has a humid temperate climate (Cfa) with rain all year round and hot summer. During the winter, the invasion of subpolar cold air is frequent, through the south winds, which makes the air temperature drop considerably, producing, in some cases, frosts.

Being a little south of the tropic of Capricorn, the sun on its way approaches the zenith only once a year, this makes the solar radiation regime present a maximum in the summer and a minimum in the winter. This is directly reflected in the temperature regime that consequently presents a simple wave during the year, that is, a maximum and a minimum. Summer is considered to include the months of December, January and February, autumn the months of March, April and May, winter the months of June, July and August and spring the months of September, October and November.

1.2. Electricity consumption data

The dataset spans from January 2017 to December 2020, and contains data generated by 55 feeders distributed in 14 substations. The number of feeders per substation is shown in Table 1.

The geographical location of each substation is presented in Table 2. Column “Code” refers to the identifier associated to each substation in the dataset. This information is also available in the substation-geographical-location.csv file in the repository.

The dataset includes a total of 1,848,947 records, generated by approximately 130,000 residential and 30,000 non residential clients. As a reference, in 2019 this region had a consumption of approximately 1000 GWh, of which 60% were residential and 40% non-residential.

The output electrical current comes from the feeder meter, therefore, the measure includes the consumers electricity consumption and the distribution losses. There is no distributed renewable generation to consider in the area.

Fig. 2 shows a sample data of the raw dataset. The complete data can be found in electricity-consumption-raw.csv file included in the repository. Table 3 shows a brief description of variables and metric unit used. As it can be noticed from Fig. 2, the first row is the header specifying the variable names, while the data starts from the second row.
datetime, substation, feeder, consumption
2017-01-01T00:00:00, A, A1,
2017-01-01T01:00:00, A, A1, 58
2017-01-01T02:00:00, A, A1, 58
2017-01-01T03:00:00, A, A1, 58
2017-01-01T04:00:00, A, A1, 52

Fig. 2. Electricity consumption: sample data.

Table 3
Electricity consumption set: description of available variables.

| ID | Feature     | Description                                                                 |
|----|-------------|-----------------------------------------------------------------------------|
| 1  | Datetime    | Date and time when measure was registered. Format: ISO 8601, yyyy-mm-ddTHH:MM:SS. |
| 2  | Substation  | Part of an electrical generation, transmission, and distribution system where the voltage is transformed from high to low, or the reverse. |
| 3  | Feeder      | Power lines through which electricity is transmitted in power systems. Electric energy meters are installed in each feeder. |
| 4  | Consumption | Electricity consumption in amperage. |

Fig. 3. Combo bar chart representing the percentage and total numbers of missing consumption data on each feeder.

1.3. Electricity consumption, missing data

Missing or null values can be found in the data, as it can be noticed in the second row of Fig. 2, where the value of the variable “consumption” is missing. A ranking of feeders, according to the number of missing values, are represented in Fig. 3. Feeders are arranged according to the number of missing records, from left to right. We can notice that the feeder labelled as I1 is the one presenting the higher number of missing values. It has almost 9% of missing records, which corresponds to about 3,000 missing values (right y-axis).

In Fig. 4, we show an hourly representation of the missing values for feeder I1. We can notice that the hours with higher missing data are those between the midnight and the first hours of the morning. On the other side, the feeder with less missing values is E2, with less than 375 missing records corresponding to 1%.

1.4. Electricity consumption analysis

We can perform an analysis of consumption curves in order to detect differences or similarities between feeders. For example, a comparison between feeders K3 and H2, shows differences
on their consumption curves during each day of week, see Figs. 5 and 6. However, they are similar regarding the distribution of the amounts of current consumption, see Fig. 7.

Fig. 5 shows H2 increases consumption around 5 AM and decreases around 5 PM from Monday to Saturday, while on Sundays it has a low and stable consumption which tell us that it is probably a feeder that serves a commercial area because usually companies and shops start their activities early in the morning and close at sunset. On the other hand Fig. 6 shows that K3 has a very stable consumption for all days of week and it has a peak around 5 PM. This may suggest that K3 is probably a feeder for a residential area because on a normal day, people go out of their work around that time and switch on their lights and devices in their houses.

Fig. 7 shows the consumption distribution of each feeder. Data are depicted through their quartiles using box plot. It can be noticed that feeders B4, E2, G4 and N1 have a higher average consumption than the others, while feeder F2 has the lowest consumption.

1.5. Meteorological data

The meteorological data spans from January 2017 to December 2020 as well, consists of a total of 22,445 records of temperature, relative humidity, wind speed and atmospheric pressure registered at the station of Minga Guazú airport located at latitude −25.45374049147125 and longitude −54.8432164928176. Fig. 8 shows a sample of the raw data. The entire dataset can be found in the meteorological-raw.csv file included in the repository. Table 4 describes the
Fig. 6. Combo bar chart representing electricity consumption registered on feeder K3 per day of week.

Fig. 7. Electricity consumption distribution of all feeders.

datetime, temperature, humidity, wind_speed, pressure
2017-01-01T00:00:00, 26, 85, 9.3, 982.5
2017-01-01T03:00:00, 25, 94, 7.4, 981.8
2017-01-01T06:00:00, 22, 92, 14.8, 981.3
2017-01-01T09:00:00, 25.2, 90, 18.5, 983.3
2017-01-01T12:00:00, 28.8, 79, 12.1, 985.3

Fig. 8. Meteorological set: sample data.

Table 4
Meteorological set: description of available variables.

| ID | Feature          | Description                                                                 |
|----|------------------|------------------------------------------------------------------------------|
| 1  | Datetime         | Date and time when measure was registered. ISO 8601 format: yyyy-mm-ddTHH:MM:SS. |
| 2  | Temperature      | Temperature in celsius.                                                      |
| 3  | Humidity         | Relative humidity in percentage.                                            |
| 4  | Wind Speed       | Wind speed in kilometers per hour (km/h).                                   |
| 5  | Pressure         | Atmospheric pressure at the station level in hectopascal (hPa).              |
variables and metric unit used. As it can be noticed from Fig. 8, the first row is the header specifying the variable names, while the data starts from the second row.

This data did not require much preprocessing, since it presented just some few missing values. Table 5 shows the number of missing values for each feature. The feature with the most missing records is the relative humidity, which has eighty one missing values, or, in other words, 0.36% of the whole set. We used a spline interpolation in order to fill the missing values. The resulted set can be found in the repository with the name: meteorological-processed.csv.

2. Experiment, Materials and Methods

Electricity consumption and meteorological data were extracted from spreadsheets provided by the Administración Nacional de Electricidad (ANDE) and the Dirección Nacional de Aeronáutica Civil del Paraguay (DINAC), respectively. The data was anonimized by replacing the original names of the feeders.

2.1. Outliers detection and data imputation of the electricity consumption data

In order to detect outliers, we used the algorithm proposed by Vallis et al. [9]. This algorithm implements the “Seasonal and Trend decomposition using Loess” (STL) [3] to obtain, given a time series $X$, the components of seasonality $S_x$, trend $T_x$ and remainder $R_x$, such that $X = S_x + T_x + R_x$. This decomposition method allows the seasonal component to be varied according to the nature of the series and is robust to the presence of outliers in the time series.

Before performing the decomposition of the time series, we have also resorted to the Box–Cox transformation technique [1] to stabilize the variance in the data so that it remains stationary and obtain an additive time series as described by Chatfiel [2] and Hyndman et al. [6]. The Box–Cox transformation is expressed as follows:

$$X^* = \begin{cases} \frac{X^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log X, & \text{if } \lambda = 0 \end{cases}$$

where $X^*$ represents the transformed time series and $\lambda$ represents the transformation parameter obtained by the maximum likelihood estimation such that data it closely resembles a normal distribution. Therefore, given the transformed data, we proceeded to apply the STL decomposition and recalculate the remainder component as $R_{X^*} = X^* - S_{X^*} - \bar{X}^*$, where $\bar{X}^*$ is the median of the data considering a non-overlapping moving window with 2 weeks length as described in Vallis et al.. Then, the generalized Extreme Studentized Deviate (ESD) test [8] is applied over the resulted remainder component using both median and median absolute deviation in test statistic to detect the outliers. Continuing with the preprocessing step, the seasonally adjusted data were determined and the outliers were discarded from it, which were filled in with a linear interpolation before adding the seasonal component and obtaining the time series $X^*$ without outliers [4]. Then, the inverse Box–Cox transformation was performed to obtain the original time series $X$. Finally, we eliminated the interpolated values at the beginning to fill in the initial missing data and a more appropriate imputation technique was used for filling several consecutive hours.

We carried out a second analysis on the data about outliers. The ranking of feeders according to the number of outliers detected is shown in Fig. 9. We can notice that feeder B3 is the one
Fig. 9. Combo bar chart representing the percentage and total numbers of consumption outliers detected on each feeder.

Fig. 10. Combo bar chart representing the percentage and total numbers of outliers detect per hour of the feeder B3. Presenting the highest number of outliers. It has almost 14% of outliers, which represents around 5,000 values (right y-axis).

Fig. 10 shows an hourly representation of the detected outliers on feeder B3. We can notice a higher presence of outliers between the midday and 4 PM.

Once all outliers and unwanted records were discarded, the historical average data imputation technique [7] was applied to estimate each missing record \( y_i \) as an average of \( N_H \) representative historical records \( y_j, j \in \mathcal{H}, | \mathcal{H} | = N_H \).

\[
\hat{y}_i^{HA} = \frac{1}{N_H} \sum_{j \in \mathcal{H}} y_j
\]  

(1)

The set \( \mathcal{H} \) includes all historical records whose day of the week (DOW) are the same as the missing one and within selected spans of the missing record. The DOW guarantees that the historical mean is calculated over records of same days of the week and similar seasonal characteristics. For this paper, the selected DOW span was \( \pm 6 \) weeks.

After applying the preprocessing techniques, 8 of the 55 feeders were still presenting missing values, as shown in Table 6. The presence of missing data in the time series after the application of the historical average method is due to the unavailability of data in the window under consideration for the data imputation technique in the first and last weeks of the dataset. Therefore, we proceeded to eliminate the days where there were still missing values from all the feeders, in such a way that they all have the same number of records. In total 33 days were discarded, thus the final set consists of 1428 days. The resulting data set contains the same number of variables as the raw one and can be found in the repository: electricity-consumption-processed.csv.

The source code to perform the outliers and data imputation procedure can also be downloaded from the repository: data-imputation.py.
Table 6
Missing values after preprocessing.

| Feeder | Amount of missing values | Number of days where they appear |
|--------|--------------------------|----------------------------------|
| C1     | 397                      | 29                               |
| C2     | 393                      | 28                               |
| F1     | 335                      | 14                               |
| F2     | 335                      | 14                               |
| I1     | 12                       | 2                                |
| I4     | 7                        | 2                                |
| L4     | 361                      | 16                               |
| N4     | 361                      | 16                               |

**Ethics Statement**

The authors declare no ethical conflicts with raw data collected and presented in this article.

**Declaration of Competing Interest**

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

**CRediT Author Statement**

**Gustavo Velázquez**: Investigation, Visualization, Writing – original draft; **Félix Morales**: Investigation, Visualization; **Miguel García-Torres**: Conceptualization, Validation, Writing – review & editing, Funding acquisition, Project administration; **Francisco Gómez-Vela**: Conceptualization, Validation, Writing – review & editing; **Federico Divina**: Conceptualization, Validation, Writing – review & editing; **José Luis Vázquez Noguera**: Resources, Writing – review & editing, Funding acquisition; **Federico Daumas-Ladouce**: Visualization, Writing – review & editing; **Carlos Sauer Ayala**: Validation, Writing – review & editing; **Diego P. Pinto-Roa**: Writing – review & editing; **Pedro E. Gardel-Sotomayor**: Data curation, Writing – original draft, Writing – review & editing; **Julio César Mello Román**: Data curation, Writing – original draft, Writing – review & editing.

**Acknowledgments**

This research was funded by CONACYT, Paraguay, grant number PINV18-661. Funding for open access publishing: Universidad Pablo de Olavide/CBUA.

**References**

[1] G.E. Box, D.R. Cox, An analysis of transformations, J. R. Stat. Soc. 26 (2) (1964) 211–243.
[2] C. Chatfield, The Analysis of Time Series: An Introduction, Chapman and Hall/CRC, 2003.
[3] R.B. Cleveland, W.S. Cleveland, J.E. McRae, I. Terpenning, Stl: a seasonal-trend decomposition, J. Off. Stat. 6 (1) (1990) 3–73.
[4] M. Friedman, The interpolation of time series by related series, J. Am. Stat. Assoc. 57 (300) (1962) 729–757.
[5] B. Grassi, F. Vázquez, R. Rodríguez, Evidencias científicas e impactos económicos del cambio climático en el departamento de alto paraná. Mades-stp. Asunción, Paraguay. (2020).
[6] R.J. Hyndman, G. Athanasopoulos, Forecasting: Principles and Practice, OTexts, 2018.
[7] J. Peppanen, X. Zhang, S. Grijalva, M.J. Reno, Handling bad or missing smart meter data through advanced data imputation, in: 2016 IEEE Power Energy Society Innovative Smart Grid Technologies Conference (ISGT), 2016, pp. 1–5, doi:10.1109/ISGT.2016.7781213.
[8] B. Rosner, Percentage points for a generalized ESD many-outlier procedure, Technometrics 25 (2) (1983) 165–172.
[9] O. Vallis, J. Hochenbaum, A. Kejariwal, A novel technique for long-term anomaly detection in the cloud, 6th [USENIX] Workshop on Hot Topics in Cloud Computing (HotCloud 14), 2014.