Data in Brief 54 (2024) 110452

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

A comprehensive dataset integrating household energy consumption and weather conditions in a north-eastern Mexican urban city

Baldemar Aguirre-Fraire, Jessica Beltrán, Valeria Soto-Mendoza*

Centro de Investigación en Matemáticas Aplicadas, Universidad Autonoma de Coahuila, Mexico

ARTICLE INFO

Article history:
Received 1 February 2024
Revised 12 April 2024
Accepted 15 April 2024
Available online 18 April 2024

Dataset link: Household energy consumption enriched with weather data in northeast of Mexico (Original data)

Keywords:
Empirical data collection
Machine learning
Forecast
Artificial intelligence
Electricity consumption behaviour
Smart plug
Environmental sensing
Time series

ABSTRACT

The prediction of domestic electricity consumption is relevant because it helps to plan energy production, among many other benefits. In this work a dataset was collected from one house in an urban city of north-east of Mexico. An ad-hoc acquisition system was implemented to collect the data using a smart meter and the open weather API. The data was collected every minute over a period of 14 months since November 5, 2022, to January 5, 2024. The dataset contains 605,260 samples of 19 variables related with energy consumption and weather data. This dataset is specifically tailored for predicting domestic energy consumption and understanding consumption behaviours, filling a void in the existing literature where such datasets for Mexico are scarce. Moreover, the multivariate nature of the dataset allows researchers to investigate and propose new techniques for forecasting or pattern classification using multivariate data collected in a real scenario.

* Corresponding author.
E-mail address: vsoto@uadec.edu.mx (V. Soto-Mendoza).

https://doi.org/10.1016/j.dib.2024.110452

Specifications Table

Subject	Energy
Specific subject area	Household energy consumption and weather data
Type of data	Raw, CSV file
Data collection	Real-time energy consumption data was collected using a smart meter device model AT-Q-SY1 connected at the main source of a domestic house over a period of 14 months. In addition, we used the free version of OpenWeather API to collect meteorological data specific to the residence's geographic location. The acquisition system includes communication with the smart meter through WiFi using a Raspberry Pi 4 microcomputer which also consumes data from the OpenWeather API. Three new energy consumption variables were computed and the whole data was structured into a single vector. The data was stored in a MongoDB Atlas cloud cluster.
Data source location	An urban city in the north-east of Mexico
Data accessibility	Repository name: Household energy consumption enriched with weather data in north-east of Mexico Data identification number: 10.17632/tvhygj8rgg.1 Direct URL to data: https://data.mendeley.com/datasets/tvhygj8rgg/1

1. Value of the Data

- The dataset contributes to the academic community to develop, train and evaluate models designed to forecast household energy consumption and understand patterns of consumption behaviours. Since the dataset provides minute-by-minute datapoints over a 14-month period, it allows to researchers to perform comprehensive analysis and to aggregate data at their convenience according to their investigation goals. In addition, the timestamps on each datapoint allow to extract valuable information, such as weekends and holidays related with consumption patterns, which can be integrated in Machine Learning models. Moreover, the dataset contains multivariate the data, including energy and weather data variables, which can be used to design more powerful models compared with univariate datasets.
- The dataset is useful for private and governmental entities in charge of generate and distribute energy to end users. For instance, the Federal Electricity Commission (CFE in spanish) in México distributes energy to the general population and the dataset allow to exploit Machine Learning techniques to predict consumptions and to understand Mexican people energy consumption patterns. These entities can use predictions to generate the required energy without waste or risk of undersupply and to plan distribution strategies. It is worth mentioning that these types of datasets with energy consumption information in real settings are scarce in México.
- Since the dataset encompasses both energy consumption and weather data, the outcomes derived from the dataset and the designed Machine Learning techniques can be used by government authorities in México to make decisions about the planning for energy transition to cleaner energy sources. This in turn, is valuable to follow global standards policies and international agreements, such as the USMCA, between North American countries.
- The dataset is also valuable to individuals, since the outcomes are informative about their own consumption patterns, with the possibility of inducing a change in their behaviour towards more awareness.

2. Background

The motivation to construct this dataset is to develop and evaluate techniques for predicting energy consumption and understand the consumption patterns in a household [1]. This is useful since an accurate prediction not only streamline energy production management, minimizing waste and optimizing resources but also promote awareness and control among consumers

regarding their energy usage, impacting both the environment and their electricity bills. Incorporating insights from previous authors [2,3], who advocate for the inclusion of climate variables to enhance model predictions, we have enhanced our dataset to encompass critical information like temperature and humidity. These factors influence energy consumption patterns, particularly in regions with extreme weather conditions such as the north-eastern part of Mexico, where the use of air conditioning or electrical heaters is more prevalent and might be different from other regions of the world [4]. The insights derived from an energy consumption dataset contribute significantly to the advancement of community prediction. These findings serve as valuable information for informing both service providers and governmental bodies about consumption trends. This, in turn, facilitates the provision of high-quality services and the development of strategic plans for transitioning to cleaner and more sustainable energy sources.

3. Data Description

The dataset is a CSV format file that contains 605,260 samples (rows) with the 19 variables (columns) described in Table 1. The Source indicates where data comes from, the Variable is the name of the dataset column, a brief Description is included, the Data type of the variable, and the Unit of each variable as well [5].

Fig. 1 depicts the behaviour of the time serie for active_power, current, temp and humidity variables for the period from November 5 to December 4, 2022.

Table 1
Detailed description of the dataset variables. Source: Own elaboration.

Source	Variable	Description	Data type	Unit
Metadata	*date*	Day and hour when the sample was taken	Date time	YYYY-mm-dd HH:mm
Smart meter	*active_power*	Active power (P)	Float	Watt
	current	Electric current (I)	Float	Ampere
	voltage	Electric tension (V)	Float	Volt
	apparent_power	Computed apparent power (S) $S = V * I$	Float	Volt-ampere
	reactive_power	Computed reactive power (Q) $Q = \sqrt{S^2 - P^2}$	Float	Volt-ampere reactive
	power_factor	Computed power factor (PF) $PF = {}^{P}/_{S}$	Float	–
OpenWeather	*main*	General description of weather conditions (*Clear, Mist, Clouds, Haze, Fog, Drizzle, Rain, Thunderstorm, Dust*)	Categorical	–
	description	Detailed description of weather conditions (*clear sky, mist, few clouds, scattered clouds, broken clouds, overcast clouds, haze, fog, light intensity drizzle, light rain, moderate rain, thunderstorm, heavy intensity rain, very heavy rain, thunderstorm with rain, drizzle, dust*)	Categorical	–
	temp	Temperature	F loat	°C
	feels_like	Temperature sensation	F loat	°C
	*temp*_min	Minimum temperature	F loat	°C
	*temp*_max	Maximum temperature	F loat	°C
	pressure	Atmospheric pressure	Integer	ATM
	humidity	Humidity percentage	Integer	%
	speed	Wind speed	Float	Km/h
	deg	Wind direction	Integer	°
	temp_t + 1	Forecasted temperature	F loat	°C
	feels_like_t + 1	Forecasted temperature sensation	F loat	°C

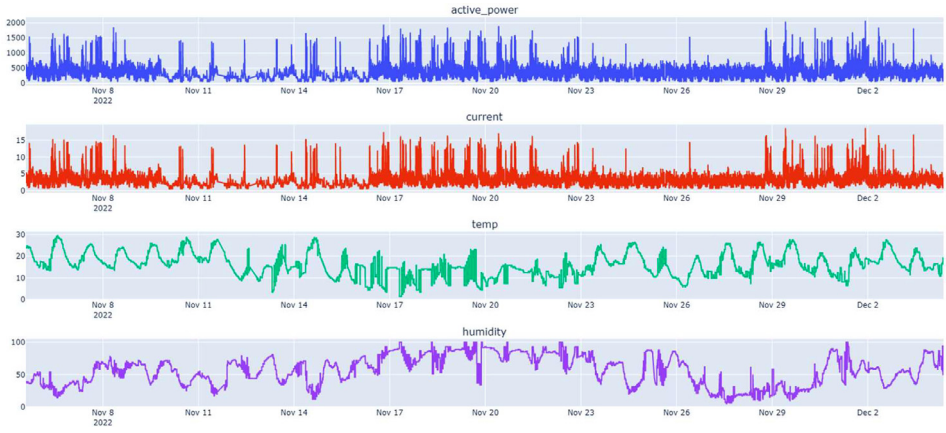

Fig. 1. Time series behaviour of active_power, current, temp, and humidity variables for the period from November 5 to December 4, 2022. Source: Own elaboration.

4. Experimental Design, Materials and Methods

4.1. Location

Data was collected for 14 months from a domestic house in the north-east of Mexico, starting on November 5, 2022, to January 5, 2024. It is pertinent to note that the data collection and processing remained unaffected by time changes due to daylight saving time. This is significant as the last change in daylight saving time for Mexico occurred in October 2022.

4.2. Data acquisition system

An ad-hoc data acquisition system was designed and implemented to collect energy and weather data. The whole system is depicted in Fig. 2.

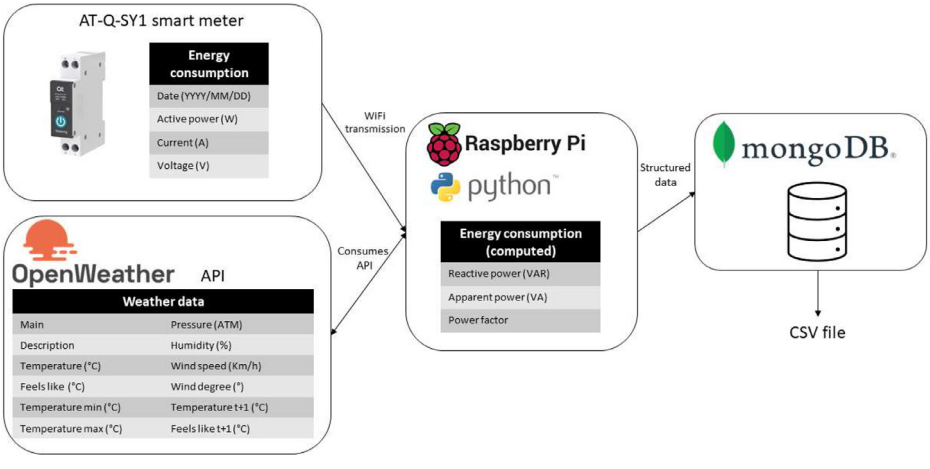

Fig. 2. Diagram of the acquisition system. Source: Own elaboration.

4.3. Energy consumption data

We acquired real-time energy consumption data using a smart meter device model AT-Q-SY1[1] connected at the main source of a domestic house. The AT-Q-SY1 WiFi smart meter is an electrical measurement device that records energy consumption at regular intervals, for this work we set it to capture samples every minute. The smart meter digitally transmits the data using WiFi communication. The device was configured to transmit energy consumption raw data to a Raspberry Pi 4[2] microcomputer where a Python script computed three new energy consumption variables.

4.4. Weather data

In addition, we used the free version of OpenWeather API[3] to collect meteorological information for the geographic location of the household using a Python script running on the same microcomputer. Both data, energy consumption and weather, were stored in a MongoDB Atlas[4] cloud cluster. Finally, a copy of the structured data was obtained from the database and saved as a CSV format file.

Limitations

This dataset presents certain limitations, primarily stemming from occasional missing data caused by disruptions in both energy and Internet services. While these gaps pose challenges, they also mirror real-world data collection scenarios, prompting researchers to explore effective pre-processing techniques to address such occurrences. Table 2 presents the statistical summary of gaps lengths.

Table 2
Statistical summary of gap lengths. Source: Own elaboration.

Statistical variable	Value
Total number of gaps	1190
Minimum gap length (minutes)	1
Maximum gap length (minutes)	1572
Mean gap length (minutes)	7.3765
Median gap length(minutes)	1
Standard deviation of gap lengths (minutes)	55.6967
1st quartile (minutes)	1
3rd quartile (minutes)	2

In Fig. 3, a scatter plot showcases the active power data for each minute, notably highlighting instances of missing data (highlighted in orange) occurring around November 12th, 2022. These gaps represent approximately 1.43 % of the total dataset.

Three duplicate records were identified in the dataset. These duplicates correspond to the index pairs: (5131,5233), (5892,5894) and (378665,378667). Duplicate records can occur due to intermittent internet connection.

Moreover, an additional constraint arises from privacy concerns, inhibiting the disclosure of information about residents' identities due to the potential inference of behaviours from the data.

[1] https://at-ele.com/.
[2] https://www.raspberrypi.com/products/raspberry-pi-4-model-b/.
[3] Weather data provided by OpenWeather, https://openweathermap.org/.
[4] https://www.mongodb.com/atlas/database.

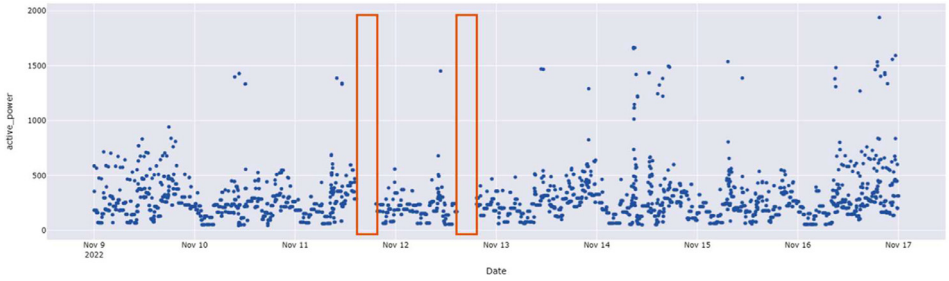

Fig. 3. Missing data of variable active_power. Source: Own elaboration.

Ethics Statement

The data acquisition system was installed in one house, informed consent was obtained from the owner and participant data has been fully anonymized. An ethical committee approved the data acquisition protocol number CIMA-CE-2022-P02 accordingly with the notice of privacy of the Universidad Autonoma de Coahuila (UAdeC).

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work the authors used ChatGPT to correct grammar and improve style, not to generate new content. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Data Availability

Household energy consumption enriched with weather data in northeast of Mexico (Original data) (Mendeley Data).

CRediT Author Statement

Baldemar Aguirre-Fraire: Conceptualization, Software, Validation, Formal analysis, Investigation, Data curation, Visualization; **Jessica Beltrán:** Conceptualization, Methodology, Formal analysis, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision, Funding acquisition; **Valeria Soto-Mendoza:** Conceptualization, Methodology, Formal analysis, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Acknowledgements

This research was partially funded by CONAHCYT through CVU 1243854 scholarship and by the Centro de Investigación en Matemáticas Aplicadas (CIMA-UAdeC).

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.

References

[1] S. Mischos, et al., Household electricity consumption in Greece: a dataset based on socio-economic features, Data Br. 48 (2023) 109232.
[2] P. Lusis, K.R. Khalilpour, L. Andrew, A. Liebman, Short-term residential load forecasting: impact of calendar effects and forecast granularity, Appl. Energy 205 (2017) 654–669.
[3] Y. Mu, M. Wang, X. Zheng, H. Gao, An improved LSTM-Seq2Seq-based forecasting method for electricity load, Front. Energy Res. 10 (2023) 1093667.
[4] B. Mashhoodi, T. Bouman, Exploring energy geography: data insights on household consumption, Data Br. (2024) 110191.
[5] B. Aguirre-Fraire, J. Beltrán, V. Soto, Household energy consumption enriched with weather data in northeast of Mexico, Mendeley Data (2024) [Online]. Available: https://data.mendeley.com/datasets/tvhygj8rgg/1.