Exploratory Data Analysis for Building Energy Meters Using Machine Learning

Analisis Eksplorasi Data Meter Energi Bangunan Menggunakan Machine Learning

Rudy Yulianto1, Sukardi2, Faqihudin3, Meika Syahbana Rusli4, Adhitio Satyo Bayangkari Karno5, Widi Hastomo*6, Nia Yuningsih7, Nada Kamilia8

1Student Doctoral of Industrial Technology Faculty, Agricultural Institute of Bogor
234Industrial Technology Faculty, Agricultural Institute of Bogor Fateta Building, Agricultural Institute of Bogor Darmaga 16002, Indonesia
5Department of Information System; Faculty of Engineering, Gunadarma University Margonda Raya Street No.100, Depok, West of Java, Indonesia
6Department of Information Technology, Ahmad Dahlan Institute of Technology and Business Ir. H Juanda Street No. 77 South of Tangerang, Banten, Indonesia
7Department of Psychology; Faculty of Psychology Science, Gunadarma University, Depok Margonda Raya Street No.100, Depok, West of Java, Indonesia
8Department of Information System; STMIK Jakarta STI&K, Jakarta BRI Radio Dalam Street No. 17, Jakarta, Indonesia

*Corresponding author: widie.has@gmail.com

Received on 01-01-2023, accepted on 25-07-203, published on 30-07-2023

Abstract

The purpose of this research was to apply exploratory data analysis techniques to building energy meters, such as electricity, cold, heat, and steam meters. A thorough understanding of energy usage patterns becomes increasingly vital in an era of growing awareness of energy management and sustainability. Trends, patterns, and anomalies can be identified in building energy meter data using meticulous data exploration approaches, which can give significant insights for increasing energy efficiency. Exploratory data analysis combined with machine learning approaches may be used to reveal hidden patterns of energy usage and examine the links between relevant factors. The findings of this exploratory data analysis gave vital insights into building energy use trends. Some significant and hidden information that was crucial for understanding energy usage within a certain time frame in each building was discovered via the investigation of the data used in this study. Steam had the highest use, whereas electricity had the lowest. Utilities were more popular before 5 a.m., followed by healthcare, with daytime use hours beginning around 10 a.m., depending on the area. During the working day, the industry needs more energy. Places of worship use more energy on weekends. There was a significant relation between the number of floors and spaces per level of a building and the height meter reading between May and October. There is a significant association between the kind of buildings used for schools, workplaces and high energy use. This study significantly contributed to the management of the energy and sustainability domains. Using exploratory data analysis and machine learning approaches to building energy meters could optimize energy usage, minimize running costs, and enhance overall energy efficiency. This research is still very open to be continued using other methods, to obtain other hidden information.

Keywords: Building, Energy Meters, Machine Learning, Data Analysis.
Abstrak

Kata kunci: Gedung, Meter Energi, Machine Learning, Analisis Data.

I. INTRODUCTION

Existing structures use around half of overall global energy consumption [1]. Since 2020, European building energy consumption laws require "structures that virtually match the NZEB (Net Zero Energy Building) requirements", particularly for new buildings [2], [3]. Smart meter infrastructure for urban structures is focused on properties and districts. While energy management for buildings is not yet accessible, measuring the energy consumed by property and structures attempts to recoup the money. To correctly spend money, we must, of course, consider energy, because the energy spent takes it straight to the end consumer [4]. Aside from expenses, energy emits carbon. A smart meter is a component of urban manufacturing and carbon impact regulation [5].

The meter is physically read once a month. Meters and data can only have an effect on one channel of transfer of energy. The smart meter analyzes energy use data in real-time and sends it to the utility remotely. They're useful for gathering data both locally and remotely, detecting long-term trends, and responding to real-time or short-term events or coordination mechanisms like management control. With modern smart meter technologies, two-way energy flow measurements are now achievable [6].

Data monitoring provides improved control over energy use as a result of networked renewable electricity, power storage, and electric cars. Because of the influence of distributed renewable power generation and battery storage, this will be crucial in future urban environments after CO2 emissions [7]. Customer perceptions of electric vehicles and trends in energy use. Over the past 25 years, in particular, in Canada and Germany, several solar system-related seasonal thermal storage test projects based on numerous collectors for the heating supply of some residential (or administrative) buildings in cold climates have been carried out [8], [9]. These projects show that this technology is possible but very expensive [10].

As a result, power electronics systems are a crucial enabler of technology for smart grids. Furthermore, smart meters serve as a link between green buildings and the energy grid. A two-way power conditioning module is necessary to achieve the interaction between the two systems [11]. Significant investments have been made in building efficiency to minimize costs and emissions. Success rates are determined by a variety of factors, including performance-based financing that pays building owners based on the difference between their current energy usage and their usage without retrofitting [12]. Improving energy adaptability can be crucial because there is no way to predict how much energy a building will use until it is repaired. So far, the counterfactual model is our best hope. Calculating savings from building renovations, energy
usage reductions compared to pre-renovation building model values, and retrofitting older technology features. A more accurate model will help increase retail encouragement and enable cheaper loans.

Several more studies similar to this research that have been undertaken can give a broader view on the application of machine learning, data mining, and energy analytics in evaluating building energy meters. Deep learning ideas and techniques are thoroughly understood. This may be used to analyze building energy data with machine learning algorithms [13]. When analyzing building energy meters, it is critical to pick the appropriate data representation. This source provides information on numerous representational learning algorithms that may be used in energy data analysis [14]. A strong grasp of the processes and generalization features of deep learning models is required to generate accurate and trustworthy models when assessing building energy meters using machine learning [15]. This review paper delves into the use of machine learning algorithms to anticipate building energy usage. It analyzes numerous algorithms and their efficacy in forecasting energy use, which may be applied to the machine-learning analysis of building energy meters [16], [17]. The purpose of this survey article is to investigate smart meter data analytics in the context of smart networks. It covers a variety of data analytics topics, such as data pretreatment, feature extraction, and predictive modeling. This survey's findings may be beneficial for interpreting data from building energy meters [18]. This review study looks at how data mining techniques may be used to analyze building energy use. It covers subjects including estimating energy usage, detecting anomalies, and increasing energy efficiency. The review sheds light on data mining techniques that may be used to examine data from building energy meters [19], [20]. This study focuses on predicting electric load, which is crucial to electricity meter analysis. It gives an overview of load forecasting methods, such as statistical and machine learning techniques. This paper's findings can be used to better identify patterns and trends in power usage data [21]. With reference to these sources, this work intends to examine building energy meters using exploratory data analysis techniques and machine learning approaches. Deep learning, representational learning, and generalization ideas and techniques can be used to improve knowledge of energy use patterns and maximize energy efficiency.

The novelty of this research lies in the combination of exploratory data analysis and machine learning techniques in the context of building energy meters. While exploratory data analysis and machine learning have been used separately in energy research, this study presents an approach that comprehensively integrates both methods to uncover patterns and relationships related to building energy consumption. Combination of Exploratory Data Analysis and Machine Learning: This study combines the strengths of exploratory data analysis, which enables the identification of patterns and trends in data, with the power of machine learning, which can generate predictive models and identify complex relationships among influential variables. This approach enriches the understanding of building energy consumption and helps optimize energy usage. Application to Different Types of Building Energy Meters: This research focuses not only on electricity meters but also includes cold, hot, and steam meters. In many cases, energy consumption data for different types of meters are available separately. By integrating data analysis from different energy meters, this study provides a more comprehensive insight into energy consumption patterns in buildings.

The data collection contains three years of meter-per-hour readings from more than a thousand buildings in various locations around the world. Before addressing the issues of various current models and technologies, this study aims to provide a realistic understanding of energy demand using building energy meter data from cold water, electric, heated water, and steam demand, and Data are taken from over 1,000 structures over a three-year period. A realistic and careful knowledge and understanding of a building's energy needs can reveal a variety of critical and hidden information. Because the data is a historical journey on the actual use of energy meters in buildings. This critical hidden information helps determine the appropriate technology model to solve the problem. Therefore, if these energy-efficient contributions can be more accurately estimated, large investors and commercial associations will be more inclined to invest in this area to improve development efficiency. Data processing is done using the Python language because this language is commonly used in programming related to machine learning [22]. The existence of libraries, pandas, and matplotlib makes the Python programming language easy to use for matrix simulations, displays attractive visual graphs, and many others [23], [24].
II. RESEARCH METHOD

The 3 datasets (train_df, weather_train_df, and building_meta_df) used in the current research are processed and extracted in this part. The effective and efficient use of energy in a structure, including cold water, electricity, gas, hot water, and steam, depends on knowing specific data and extracting hidden information. In general, the methodologies used in this study were data gathering, extraction, analysis, feature engineering, and data correlation.

A. Load the Dataset

The data was obtained from Kaggle.com, from the existing data 3 files were used, as follows: The file training.csv (Figure 1) contains 20,216,100 lines of data and the following features: Building id is a foreign important for constructing metadata, code for each meter id, electricity (0), cold water (1), steam (2), water heating (3), measurement time, and meter reading for power electric expenditure (kWh).

![train_df](image)

Figure 1. Train file

Weather data is gathered from the nearest meteorological station. The Weather.csv file (Figure 2) has 139,773 data lines and includes the following features: location identification (id-site), air temperature level for each place (Celsius), cloud conditions, humid temperature level (Celsius), rainfall depth per hour (mm), surface pressure conditions from the sea (milli-bar), wind direction (0-360), wind velocity (meters per second).

The file building meta.csv (Figure 3) has 1,449 lines of data and includes the following information: Identification of the place (id-site), foreign primary key connected with the weather state file, and building identity (id-building), foreign primary key for associated with the training file, main use indicates the main activity category indicators for the building, square feet represents the floor area of the building. Python libraries are used for developing.
Exploratory Data Analysis (EDA)

Training, building, and weather datasets were merged to produce the following analysis. We attempted to combine the training and test sets, however, the processing time for the subsequent EDA was so sluggish that it would repeat whatever cleanup/feature engineering was done on the train to the test set [25]. The total number of rows and columns in the combined data frame (combine train) is now 20,216,100. Feature Engineering expands over time as we develop new features, including month of the year, day of week timestamp, and hour of day.

C. Missing Data

The missing data is described below. Weather observations and missing structures should be checked throughout the report. To visualize the distribution of meter readings without outliers, dividing the outliers
using the formula \((1.5 \times IQR)\) turns out to produce a very skewed target variable, even for these observations within the outlier limits. After performing several transformations on meter_reading for modeling purposes, we obtain the distribution of meter data (Figure 4).

![Figure 4. Log transforming meter reading variable is necessary](image)

### D. Kind of meter and several sorts of meters

The ggplot package included with the Python programming language may be used to show several sorts of meter readings. For meter reading and meter types, there are four different types of meters that are shown in coded form, namely: Electricity (0), Coldwater (1), Steam (2), and Hot water (3). As we can see, these types of steam meters tend to have higher meter readings, while electricity tends to have the lowest energy readings (Figure 5). For Weekday and Meter readings, we take that to be expected, Sundays tend to have lower readings, but there isn’t a big difference between the days (Figure 6).

![Figure 5. Readings from four types of meters’ density logs](image)
The time of day and meter readings are significantly higher during traditional business hours and this is to be expected. Time of day is likely to be a significant predictor in future models. There are many variations of meter readings for each type of building. Meter readings for maintenance buildings (more than 225) with peak usage time during the day (15:00). For housing, the peak is around 80, with the peak usage time at night at 19:00. Manufacturing with meter readings of over 150 occurred around 5 am. Meanwhile, the lowest meter use only reaches 6 around 10:00 for places of worship. The highest usage meter for utility locations with usage before 5 am reaches 350 (Figure 7). For prime usage and meter readings, we can see that Utilities and Healthcare establishments tend to have the highest readings, while Religious Places of Worship the fewest - no doubt they are visited less often than high energy users (Figure 8).

Figure 6. Weekly readings from a daily density log

Figure 7. Hourly meter readings per day for all types of buildings.
It's important to note that different building types use energy differently on different days. For example, religious facilities typically use more energy on weekends than on weekdays, while industrial (commercial) facilities typically use more energy on weekdays (Figure 9). Intriguingly, the average meters reading per reading began to rise in May and peaked between July and October 2016 in terms of meters readings over time (Figure 10). With a few exceptions, this tendency is true for the most of building types. Manufacturing decreased during the peak times mentioned above, but Services, Technology, Utilities, and Warehouses stayed largely stable throughout the year (Figure 11).
Figure 10. Median meter readings between January 2016 and January 2017

Figure 11. For each utility building, the median meter readings from January 2016 to January 2017

E. Correlation

The amount of storeys and the size of the structure both have a strong Pearson correlation with meter readings, implying that the larger the structure, the greater the reading. The year of construction also has a substantial positive link - older structures may have higher values. Any model created with extremely weak negative Pearson correlations may need wind speed, air temperature, and cloudy weather (Figure 12).
Analyzing these variables will also assist us in determining how they are distributed and whether or not there are any problems with the data (missing variables, skewness, extreme outliers, etc.). There are no missing records in the primary usage variable. The educational building type is by far the most read, followed by the office building type, which is somewhat less read but still more than twice the third most read Entertainment building type. The tail is fairly long, with 11 underparts that are slightly lower (Figure 13).

The square feet variable has 0 missing values. With a median of 72,709 feet and a mean of 107,783 square feet, the square foot distribution is favorably skewed (Figure 14). The meter variable has 0 missing records.
We can see that electricity was the most often read, with a record of almost 12m, while hot water had a record of 1.26m (Figure 15).

The year built variable has 12,127,645 missing records. These years are depicted below. 11 of the 16 distinct types of service buildings had no year built at all, while 11 of the 16 different types were missing by more than half (Figure 16). With 55 structures completed, 1976 was the most often constructed year in recorded history (Figure 17). The floor count variable has 16,709,167 missing records. 1094 of the 1449 distinct building IDs do not have a floor count. This number indicates that there aren’t many buildings with more than ten levels, whereas the average number of floors is three (Figure 18).
Air Temperature. There are 96,658 missing air_temperature records in our combined training set, however as helpfully pointed out by PC Jimmy in the comments section, there are only 55 missing air temp records in the weather_train data set. I will create a function to fill these missing values.

Filling these missing values will not be a simple mean/median imputation, but rather a two step process:

a. Take the air_temperature value for the last timestamp as the next timestamps value
b. If the previous timestamp is also NA, then take the average of that day's temperature for that particular site_id

Following this process results in no further NAs in the air temperature variable. Conveniently, the function can also be applied to dew temperature.

Plotting the variable shows the variable to be fairly normally distributed, with the majority of recordings being between ~13 and 25 degrees (Figure 19). Dew Temperature, in our combined training set, there are over 100,000 missing dew temperature data. The function created above can be applied to fill these missing values (Figure 20).
Energy efficiency must be the ultimate objective in order to achieve sustainable development goals. Because of the differences in height based on location and time, the management of energy consumption looks to be quite difficult. The study's conclusions are based on relevant data analysis, which demonstrates that various factors must be addressed for the management of energy management in a building in order to make it more efficient. The experiment's findings can be summed up as follows:

a. According to the meter reading, steam has the largest consumption, while electricity has the lowest.

b. Utility is highest before 5 a.m., followed by healthcare with daytime utilization hours after 10 a.m., depending on region.

c. The industry will need more energy during working days.

d. On weekends, places of worship consume more energy.

e. Between the months of May and October, there is a considerable association between the number of storeys and the space of each level of the building and high meter readings.

f. There is a substantial association between the kind of building used for school and offices and high energy meter consumption.

Generating wiser and more knowledgeable decisions can be aided by the insights obtained from the EDA process. While improving productivity and process efficiency might present challenges and opportunities, doing so ultimately benefits the organisation. In the future, these insights will be used to sophisticated data analytics modeling of energy consumption and resource usage in data centers for energy efficiency management.

III. CONCLUSION
AUTHOR CONTRIBUTION

Conceptualization Rudy Yulianto, and Meika Syahbana Rusli.; methodology, Sukardi and Faqihuddin.; software, Adhisto Satyo Bayangkari Karno. and Widi Hastomo; writing-original draft, Rudy Yulianto.; visualization, Nia Yuningsih. and Nada Kamila.

REFERENCES


