SEP799 – Final Project Report

Title: Forecasting of Electric Vehicle Charging Loads on Distribution Systems

Community Partner: Burlington Hydro Inc.

Faculty Lead: Dr. Marjan Alavi

Group Members (Group # 3):
Asma Ahmed (400544917)
Jieming Yin (400184665)
Junting Ye (400552280)
Yinghua Ma (400552029)

Table of Contents

1. Abstract

This report presents a comprehensive study on forecasting electric vehicle (EV) charging loads on distribution systems for Burlington Hydro Inc. The growing integration of EVs into the power grid introduces significant challenges, particularly in accurately predicting load demands. By employing advanced data analysis techniques like Exponential Moving Average (EMA) and Wavelet Decomposition, alongside machine learning algorithms such as Random Forest, Logistic Regression, and Clustering, this study identifies EV charging patterns, forecasts demand on the electrical grid, and aims to assesses impacts on local transformer infrastructure. The methodology includes data analysis using Python and sophisticated optimization models like Mixed-Integer Quadratic Programming (MIQP). This report aims to enhance grid reliability, optimize resource allocation, and support sustainable urban growth through a robust and accessible EV charging infrastructure.

2. Project Description

The project's primary goal is to forecast EV charging loads to support Burlington Hydro Inc. in managing electrical load distribution efficiently. The report delves into methodologies that utilize data analysis, machine learning models, and real-time data acquisition to predict EV charging demand accurately. By understanding EV adoption patterns, charging behaviors, and the spatial distribution of residential EV chargers, the project aims to optimize resource allocation and enhance grid stability. The initiative also focuses on environmental stewardship by reducing greenhouse gas emissions and supporting the transition to a low-carbon economy.

2.1. Project Overview

Burlington Hydro Inc. (BHI) serves approximately 68,500 customers through a network of distribution lines and substations. With the increase in EV registrations, BHI faces challenges in forecasting load demands accurately, which complicates resource allocation and grid management. This project involves refining BHI's forecasting methodologies using advanced analytics and machine learning techniques. The project synthesizes data from various sources, including smart meters and demographic trends, to construct a comprehensive view of demand drivers. This approach ensures a resilient power grid capable of adapting to urban mobility and energy consumption shifts.

2.2. Objectives

The primary objective of this project is to develop a robust methodology to accurately detect EV charging events from power consumption data. This involves using EMA and Wavelet Decomposition to identify significant power usage deviations and MIQP for optimization. Clustering algorithms were used to categorize consumption patterns and enhance the detection of charging events. Additionally, the secondary objective focuses on identifying non-charging periods and detecting anomalous consumption patterns using machine learning techniques. To support these goals, the project includes several additional objectives. First, it aims to enhance the dataset with relevant features to improve the accuracy and reliability of the analysis. Second, it will experiment with different models and techniques to optimize the detection of EV charging events. Finally, the project will visualize the results to provide clear and actionable insights into power consumption and charging activities.

3. Background

Last term, we developed a detailed methodology for incorporating electric vehicle (EV) charging into the power distribution frameworks of Local Distribution Companies such as Burlington Hydro. The methodology for integrating electric vehicle (EV) charging into the power distribution networks emphasizes a three-part strategy focusing on data analysis, demand forecasting, and infrastructure assessment. The approach starts with the meticulous collection and analysis of detailed metering data to identify EV charging patterns. This is achieved using Python programming and its robust libraries, such as pandas for data management and matplotlib for data visualization, which help in dissecting complex datasets and presenting the findings in an understandable format. Tools like PyCharm and Spyder enhance the analysis process, supporting the effective handling and visualization of data. Additionally, machine learning algorithms like Random Forest and Logistic Regression were implemented to predict the presence of EV chargers and assess their grid impacts.

4. Methodology

4.1. Data Pre-processing

4.1.1. Converting XLSX File into Parquet File

The dataset provided by the community partner is in the XLSX format, each file approximately 600 MB and there are 12 files in total, and the file are password encrypted. The XLSX format, while popular due to its accessibility in software like Microsoft Excel, is not optimized for large-scale data processing and machine learning project. This is due to its row-based storage which can significantly slow down read operations, especially with large datasets. Parquet is a columnar storage file format that is optimized for use with large datasets. It allows for efficient data compression and encoding schemes, making it ideal for handling big data use cases. Converting into parquet file not only reduces the file size but also improve the speed of data loading.

4.1.2. Filter Out Irrelevant Categories

This step involves removing data that does not contribute to the analysis, ensuring only relevant information is processed.

4.2. Semi-supervised Learning

Semi-supervised learning is a machine learning approach that makes use of both labeled and unlabeled data for training. It's suitable for our dataset since there are only 10% of the data is labeled and 90% of the data is unlabeled. This technique means to train a model on a small amount of labeled data and then using the model to predict labels for the unlabeled data. (IBM, What is semi-supervised learning?, 2023)

4.2.1. Neural Network

Neural networks are well-suited for semi-supervised learning due to their ability to learn complex patterns. The chosen model is a 4-layer neural network. The training process involves:

- **Pre-training on Labeled Data**: Establishing baseline learning by training on labeled data.

- **Incorporating Unlabeled Data**: Gradually introducing unlabeled data using pseudo-labeling (assigning labels to unlabeled data based on model predictions).
- **Iterative Refinement**: Alternately updating the model using labeled data and refined pseudo-labels, which helps the model generalize better beyond the labeled dataset.

```python
# Define and train the initial model
model = Sequential([
    Dense(128, activation='relu', input_shape=(X_train.shape[1],), kernel_regularizer=l2(0.01)),
    Dropout(0.5),
    Dense(64, activation='relu', kernel_regularizer=l2(0.01)),
    Dropout(0.5),
    Dense(32, activation='relu', kernel_regularizer=l2(0.01)),
    Dropout(0.5),
    Dense(1, activation='linear')  # Output layer
])
model.compile(optimizer='adam', loss='mean_squared_error')
model.fit(X_train, y_train, epochs=20, batch_size=32, validation_data=(X_val, y_val))
```

Figure 1: Code snippet of the model

4.3. Handling Bias Dataset

The primary issue we encountered is a biased dataset. As illustrated in the diagram, 90% of the labeled data is categorized as "1." Consequently, the model predominantly learns the pattern of "1" and tends to produce predictions favoring "1."

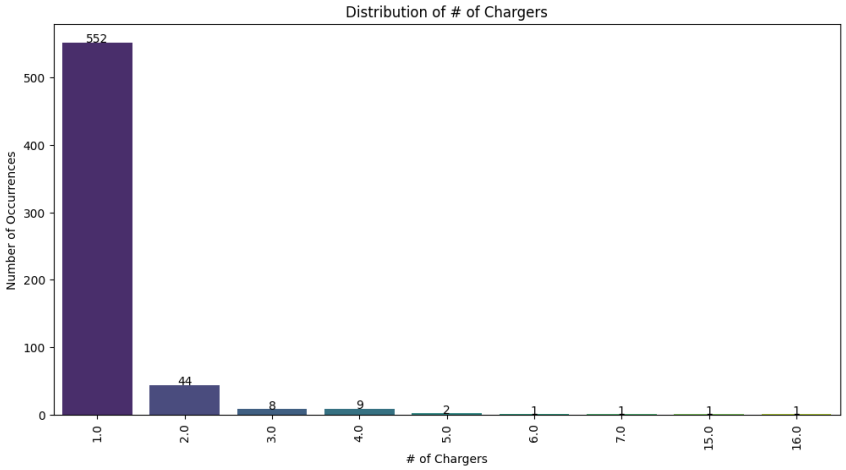


Figure 2: Distribution of the number of EV chargers

4.4. Mixed-Integer Quadratic Programming (MIQP)

MIQP is an optimization technique that involves solving problems with both integer and continuous variables, where the objective function is quadratic, and the constraints are linear. It combines elements of linear programming and integer programming, making it suitable for complex optimization problems. (IBM, MIQP: Mixed integer programs with quadratic terms in the objective function. Retrieved, 2021)

The MIQP model was employed to detect charging events by solving an optimization problem with both binary and continuous variables. The objective function aimed to minimize the absolute differences of the

baseload power, while constraints ensured the correct representation of observed power consumption and binary transitions for charging events. (Feng Li, 2024) (IBM, MIQP: Mixed integer programs with quadratic terms in the objective function. Retrieved, 2021)

Objective Function: *Minimize* $\sum Abs\_diff$

Constraints:

- Observed power equals baseload plus EV charging power: $Pobst = PBLt + PEVt$
- Link EV charging power to binary indicators: $PEVt \leq xt \times M$
- Ensure binary transitions for start and end of charging events: $y_t \geq x_t - x_{t-1}$ and $z_t \geq x_{t-1} - x_t$.

Implementation:

- Utilized the PuLP library for linear programming in Python.
- Solved the MIQP model for each row in the labeled dataset to determine charging statuses.
- Filtered out values under 1.5 kW as non-charging to improve accuracy.

4.5. Isolation Forest for Outlier Detection

The primary objective of utilizing the Isolation Forest technique is to identify outliers in the power consumption data, enabling the distinction of anomalous periods from typical electric vehicle (EV) charging events. This differentiation is crucial for maintaining the reliability and efficiency of the charging infrastructure by promptly addressing irregularities. The process involved training an Isolation Forest model specifically on the charging status data. This model was chosen for its effectiveness in handling high-dimensional data and its capability to isolate anomalies. By learning the patterns of normal charging behavior, the Isolation Forest model could accurately predict outliers. These predicted outliers represent periods that deviate significantly from established charging patterns, thereby flagging potential issues such as unexpected spikes in power consumption or irregular charging sessions. The identification of these anomalies is a critical step in ensuring the robustness and reliability of the smart parking management system, allowing for proactive measures to be taken in response to unusual charging activities.

4.6. Mixed-Integer Convex Quadric Problem (MICQP)

The Mixed-Integer Convex Quadratic Problem (MICQP) is a sophisticated optimization framework that combines the elements of mixed-integer programming with convex quadratic functions. In MICQP, the objective function includes quadratic terms and is convex, ensuring that any local minimum is also a global minimum. This feature is critical in many practical applications such as portfolio management, energy system optimization, and production scheduling, where the relationships between variables are naturally quadratic and convex. Additionally, MICQP allows for both integer and continuous decision variables, providing a detailed modeling approach for problems involving both discrete and continuous decisions. Solving MICQPs involves advanced algorithms, typically leveraging branch-and-bound strategies in conjunction with convex optimization techniques to efficiently explore the solution space (Wolsey, 1999)(Pages 447-471).

The mathematical model for Mixed-Integer Conic Quadratic Programming (MICQP) to identify electric vehicle (EV) charging events can be described as follows:

Objective Function

Minimize the squared deviation of the base load from its mean across all times:

$$Minimize \sum_{i=1}^{N}\sum_{t=1}^{T}(PBL[i,t] - mean\_BL[i])^2$$

Constraints

- **Power Decomposition Constraint**: For each location $\boldsymbol{i}$ and time $\boldsymbol{t}$, the observed power $Pobs$ is the sum of the baseline power PBL and EV charging power PEV:

$$PBL[i,t] = Pobs[i,t] - PEV[i,t]$$

- **Non-Negative Power Constraint**: Both PBL and PEV must be non-negative:

$$PBL[i,t] \geq 0, PEV[i,t] \geq 0$$

- **EV Charging Power Constraint**: The EV charging power PEV is limited by the maximum potential power level, scaled by a binary variable $x[i,t]$that indicates if charging is active:

$$PEV[i,t] \leq (\sum power\ levels) \times x[i,t]$$

- **Charging Event Constraints**: Introduce binary variables $y[i,t]\ and\ z[i,t]$ for starting and ending charging events, ensuring that a maximum of three charging events can start in any given day:

$$\sum(y[i,t-k]\ for\ k\ in\ range\,(3)) \leq 3$$

Implementation

- **Solver**: The model is solved using the cvxpy library with Gurobi as the solver, accommodating the quadratic nature of the objective and the integrality constraints on some decision variables.
- **Data Handling**: Python libraries like numpy and pandas are employed for data manipulation, feeding data into the optimization model.

This formulation closely mirrors the provided MIQP model but is adapted specifically for the data and requirements of monitoring and predicting EV charging patterns based on electrical consumption data

4.7. Exponential Moving Average (EMA) and Wavelet Decomposition

The Exponential Moving Average (EMA) is a weighted moving average technique that places greater importance on more recent data points, making it effective for identifying trends in power or energy consumption. For instance, analyzing hourly power usage data can reveal spikes that indicate potential electric vehicle (EV) charging events. The EMA is calculated using the formula:

$$EMAt = \alpha \times Power\ Usaget + (1 - \alpha) \times EMAt - 1$$

where α is the smoothing factor, determining the rate at which older data points decrease in significance. By setting a deviation threshold, such as 20% above the EMA, unusual consumption patterns can be flagged in real-time. This method provides a dynamic way to monitor power usage and detect anomalies, as it continuously adapts to the latest data (Davis, 2016) (Hyndman, 2013).

Wavelet decomposition, on the other hand, breaks down a signal into its constituent components at various levels of resolution using wavelets. Wavelets are small waves that are scaled and shifted to capture both frequency and location information. One common type of wavelet used is the Daubechies wavelet ('db1'). The process involves decomposing the power usage signal into wavelet coefficients using a discrete wavelet transform (DWT), extracting approximation coefficients to represent the baseline trend, and reconstructing the signal using only these approximation coefficients. The mathematical representation of wavelet decomposition is:

$$f(t) \approx \sum j, kcj, k\psi j, k(t)$$

where $\psi j, k(t)$ are the wavelet basis functions, and cj, k are the wavelet coefficients. This method effectively isolates the underlying patterns from noise, allowing for the identification of significant deviations indicative of EV charging events (Mallat, 1999).

Both EMA and wavelet decomposition enhance the detection of EV charging behaviors by providing a robust framework for analyzing power usage data. The smooth wavelet baseline highlights significant power usage deviations, aiding in EV charging event detection. Both methods are beneficial in this case, and a combine application of both will provide a better result.

As shown in the code below, if the selected day is available, it extracts the hourly power usage. Using wavelet transforms, it decomposes the data to obtain a smoothed baseline and calculates an Exponential Moving Average (EMA) for trend detection. Spikes are identified when the power usage exceeds the EMA by a threshold of 0.5, and the neighboring hours and changing significantly. And plateaus are detected when usage remains above the wavelet baseline by a threshold of 1, for a specified duration.

```
# Setup thresholds and other constants
wattage_threshold = 2.4
excess_over_ema = 0.5
excess_over_wavelet = 1
plateau_duration = 3  # hours

if not selected_day.empty:
    hourly_data = selected_day.iloc[0, 4:28].values.astype(float)
    coeffs = pywt.wavedec(hourly_data, 'db1', level=2)
    approx = coeffs[0]
    details = coeffs[1:]

    ema_data = pd.Series(hourly_data).ewm(span=5, adjust=False).mean()
    wavelet_base = pywt.waverec([approx] + [np.zeros_like(d) for d in details], 'db1')

    spikes = []
    plateaus = []
    count = 0  # To count consecutive hours above thresholds

    for i in range(1, len(hourly_data) - 1):
        if (hourly_data[i] > ema_data[i] + excess_over_ema and
            hourly_data[i] > hourly_data[i-1] + 0.3 * hourly_data[i-1] and
            hourly_data[i] > hourly_data[i+1] + 0.3 * hourly_data[i+1] and
            hourly_data[i] > wattage_threshold):
            spikes.append(i)

        if (hourly_data[i] > wavelet_base[i] + excess_over_wavelet):
            count += 1
        else:
            if count >= plateau_duration:
                plateaus.extend(range(i-count, i))
            count = 0

    # Check if the last hours form a plateau
    if count >= plateau_duration:
        plateaus.extend(range(len(hourly_data) - count, len(hourly_data)))
```

4.8. Time Series Analysis

The method of Time Series Analysis is used for data indexed with timely or sequential order. It is widely applied in forecasting stock prices, sales figures, temperature, etc. The residential data of power consumption is well segmented and recorded in hours, days, and months, which provides a suitable basis for time series analysis.

Time Series Analysis treats the timely data with four main factors, respectively trend, seasonality, cyclicality, and randomness. Trend is the long-term change pattern in the data. Seasonality is the periodic (or seasonal) fluctuation capturing regular data changes. Cyclicality is the longer-term cyclical change compared to seasonality. Randomness is the unexpected and unpredicted random variation.

4.8.1. Model

The Seasonal Autoregressive Integrated Moving Average model (SARIMA) is used in analyzing power usage data in our experiment. It can be considered as a specially combined regression model. As its name suggests, SARIMA contains autoregression, differencing, and moving averages. The model firstly applies one-step and s-step differential (s denotes the length of seasonal periods) to the original data. Secondly, the model regresses the data onto the historical data and white noises.

4.8.2. Data Preparation

We manually selected and collected all the power usage records with registered electric vehicles. Both on a daily and hourly basis, we calculated the average power usage and denoted them as d_t and h_t. It is worth mentioning that we used independently and identically distributed white noises $\{\varepsilon_t\}$ to fit the randomness in data.

4.9. Clustering Algorithms

Clustering algorithms are a class of machine learning techniques designed to group sets of objects in such a way that objects in the same group (or cluster) are more similar to each other than to those in other groups. These algorithms can function with unlabeled data, making them particularly useful for distinguishing clusters corresponding to charging days and non-charging days in our dataset. The primary goal of clustering is to reveal the intrinsic structure of the data, which can be challenging due to subtle variations in hourly power usage data. Given the complexity of setting a threshold to determine specific charging hours, clustering provides a holistic approach to separate charging days from non-charging days without the need for explicit labels.

Clustering techniques such as K-means, hierarchical clustering, and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) are commonly used. K-means clustering, for example, partitions the data into K clusters by minimizing the variance within each cluster. Hierarchical clustering builds a tree of clusters, whereas DBSCAN groups together points that are closely packed together, marking points that lie alone in low-density regions as outliers. These techniques enable us to identify patterns and trends that might not be immediately apparent through other forms of analysis (A.K. Jain, 1999).

Because of the subtle changes in hourly power usage data, we had unprecedented difficulties in analyzing the hourly data (such as setting a threshold to determine if the user is charging the EV in this particular

hour). We decided to apply clustering algorithms to bypass this obstacle. The specific technical routes are: 1) applying clustering to divide the possible charging days and non-charging days holistically; and 2) using wavelet decomposition to filter the usage data and identify the specific times of charging.

To address the seasonal variations in power consumption, we segmented the dataset into 12 months. This approach aimed to mitigate the influence of high summer power consumption, which could skew clustering results. Clustering was performed separately for each month's data to ensure the results were relevant to the specific time of year

4.9.1. Data Preparation

Data preparation is a crucial step in the clustering process. We organized all the power usage data with registered EVs (549 unique locations * 365 days of the year 2023 * 24 hours of specific power consumption). Before applying clustering algorithms, we normalized the data by column to cancel out the influence of scale, ensuring that each feature contributes equally to the clustering process. Normalization transforms the data to a standard deviation unit for each column, which is essential for algorithms like K-means that are sensitive to the scale of the data (Jiawei Han, 2011). This preprocessing step allows the clustering algorithms to focus on the patterns within the data, leading to more accurate and meaningful clusters.

4.9.2. Initial Clustering

KMeans, Hierarchical, DBSCAN, MeanShift, GMM, Spectral, and OPTICS clustering algorithms were applied to the dataset. Each of these algorithms was selected due to its unique characteristics and ability to handle different types of data distributions and clustering challenges. To evaluate the performance of these clustering algorithms comprehensively, several key evaluation metrics were employed. These metrics included the Silhouette Score, which measures how similar an object is to its own cluster compared to other clusters; the Davies-Bouldin Index, which quantifies the average similarity ratio of each cluster with respect to its most similar one, indicating the algorithm's ability to form distinct clusters; and the Calinski-Harabasz Index, which assesses the ratio of the sum of between-cluster dispersion and within-cluster dispersion, thus providing an indication of the overall quality of the clustering. By utilizing these evaluation metrics, the effectiveness and accuracy of each clustering algorithm were thoroughly assessed, allowing for a detailed comparison, and understanding of their performance on the dataset.

4.9.3. Second Step Clustering

After the initial clustering, we performed a second step clustering with the hypothesis that charging days exhibit higher consumption than non-charging days. To substantiate this hypothesis, the second step clustering aimed to refine and verify the initial results. This step was crucial as it provided a deeper layer of analysis, distinguishing between days with high consumption typically associated with EV charging events and those without. This involved an intricate process of cross validating the clustering results with outcomes derived from Wavelet Decomposition, a technique used to break down and analyze the power usage signal into its constituent frequencies. Additionally, a Threshold Policy was applied, which set specific benchmarks for power usage to differentiate charging events from regular consumption patterns. By integrating these methods, we enhanced the confidence and accuracy in detecting actual charging days. The combination of these advanced techniques ensured that the identified charging patterns were robust and reliable, addressing potential anomalies and inconsistencies in the data, and providing a

comprehensive understanding of the EV charging behaviors. This multi-faceted approach reinforced the validity of our findings and demonstrated the effectiveness of using multiple analytical layers to achieve precise and dependable results.

5. Result and Discussion

5.1. Semi-Supervised Learning

The initial step of semi-supervised learning (training on labeled data) yielded poor results, the red line shows the result, and the blue point is our prediction. As shown in the diagram, the model predicts correctly for "value 1" and some of "value 2", however, the prediction for higher value is not correct. Therefore, further investigation is necessary before proceeding.

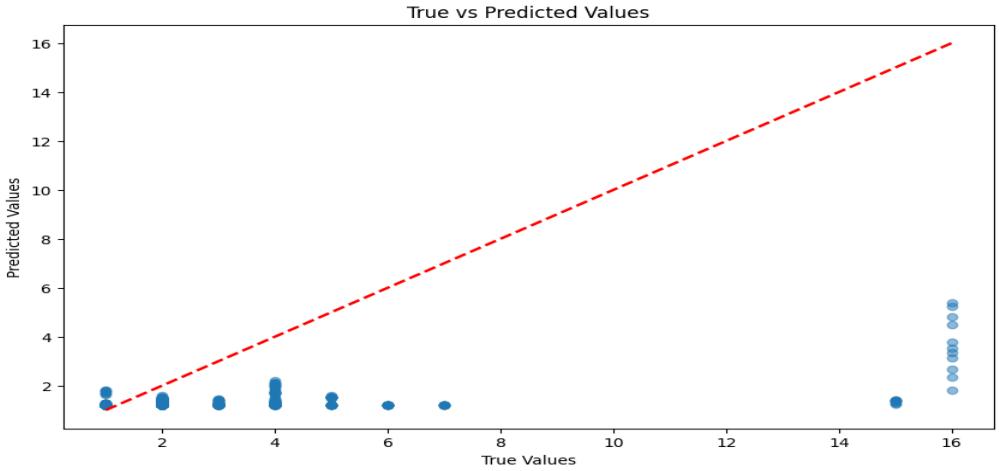


Figure 3: Semi-Supervised prediction result

5.2. Applying MIQP to the Dataset

The Mixed-Integer Quadratic Programming (MIQP) model was applied to the dataset to detect electric vehicle (EV) charging events. Initially, hourly power consumption values (R1 to R24) were extracted for each labeled data row. The MIQP model was then employed to identify charging events by generating binary charging statuses, where only power consumption values above 1.5 kW were considered as valid charging events.

The detected charging statuses were integrated back into the original dataset, resulting in the addition of new columns that represent each hour's charging status (Charging_Status_1 to Charging_Status_24). This enhanced dataset provides a comprehensive view of charging patterns, facilitating further analysis and decision-making processes related to EV charging load forecasting and management.

	YYYYMMDD	LOCATION	RATECLASS_DESC	...	Charging_Status_22	Charging_Status_23	Charging_Status_24
0	20230101	23	001: Residential: TOU	...	1	0	0
1	20230102	23	001: Residential: TOU	...	0	0	0
2	20230131	23	001: Residential: TOU	...	1	0	0
3	20230103	23	001: Residential: TOU	...	1	0	0
4	20230104	23	001: Residential: TOU	...	0	0	0

Figure 4: Sample result of EV charging status (0 or 1) using MIQP

5.3. Outlier Detection using Isolation Forest

Using the Isolation Forest for outlier detection complements the charging event detection performed by the MIQP model without affecting its accuracy or outcomes. The Isolation Forest identifies anomalies in the detected charging patterns, providing an additional layer of analysis to flag periods that deviate from typical charging behavior.

Training and Prediction

The Isolation Forest algorithm was employed with a contamination rate of 0.1 to identify outliers. This algorithm predicted outliers within the dataset, effectively highlighting non-charging periods that stood out as unusual.

Visualization

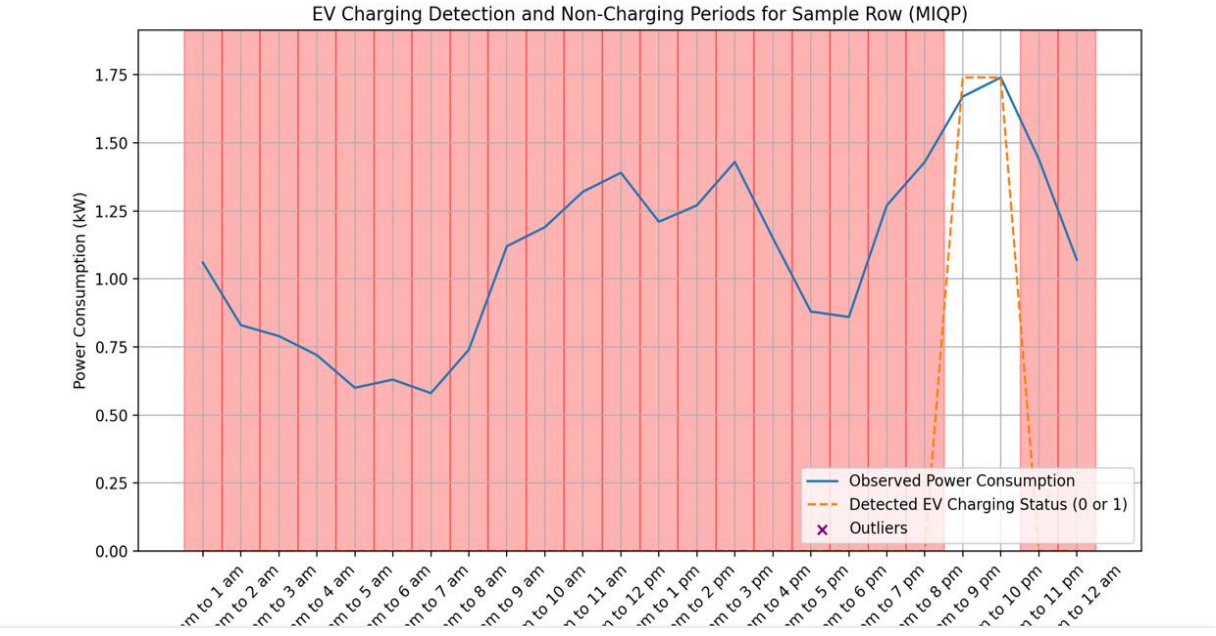


Figure 5: Detected EV Charging status using MIQP

To enhance understanding, observed power consumption in Figure 5, detected EV charging status, and identified outliers were plotted for a random sample row. Custom hour labels were used to improve readability, making it easier to interpret the relationship between normal and anomalous charging periods.

Model Performance

The performance of the model was rigorously evaluated using several metrics: accuracy, precision, recall, and F1 score. Experimentation with different values of the Big-M constant in the MIQP model demonstrated that the optimal value achieved a balance between feasibility and performance. For real-life biased data, the model exhibited promising accuracy at 60.10%, indicating its overall effectiveness in identifying true positives and negatives. However, the precision was relatively low at 9.97%, suggesting a higher rate of false positives. The recall rate stood at 37.44%, reflecting the model's capability to detect actual charging events, while the F1 score, which balances precision and recall, was 15.75%. These results highlight the model's strength in detecting a broad range of events, although there is room for improvement in precision to reduce false alarms.

5.4. Applying MICQP to the Dataset

The Mixed-Integer Convex Quadratic Programming (MICQP) model was applied to the dataset to detect electric vehicle (EV) charging events, where the power consumption values above 2.2 kW for consecutively three hours, or more were considered as valid charging events.

Updated EV Charging detection results:

	YYYYMMDD	LOCATION	EV_Charging
4028	20230501	50254	1
4029	20230502	50254	1
4030	20230531	50254	1
4031	20230503	50254	1
4032	20230504	50254	1

Figure 6: EV Charger detection result using MICQP

Figure 7 displays the observed power consumption against the MICQP-based detection status over a 24-hour period. The blue line represents the actual power usage, while the red dotted line indicates the binary detection status (0 or 1) determined by the MICQP model. The observed power consumption shows significant fluctuations, particularly during early morning and evening hours, suggesting possible EV charging events. The MICQP model successfully identifies these events, evidenced by the corresponding binary spikes in detection status.

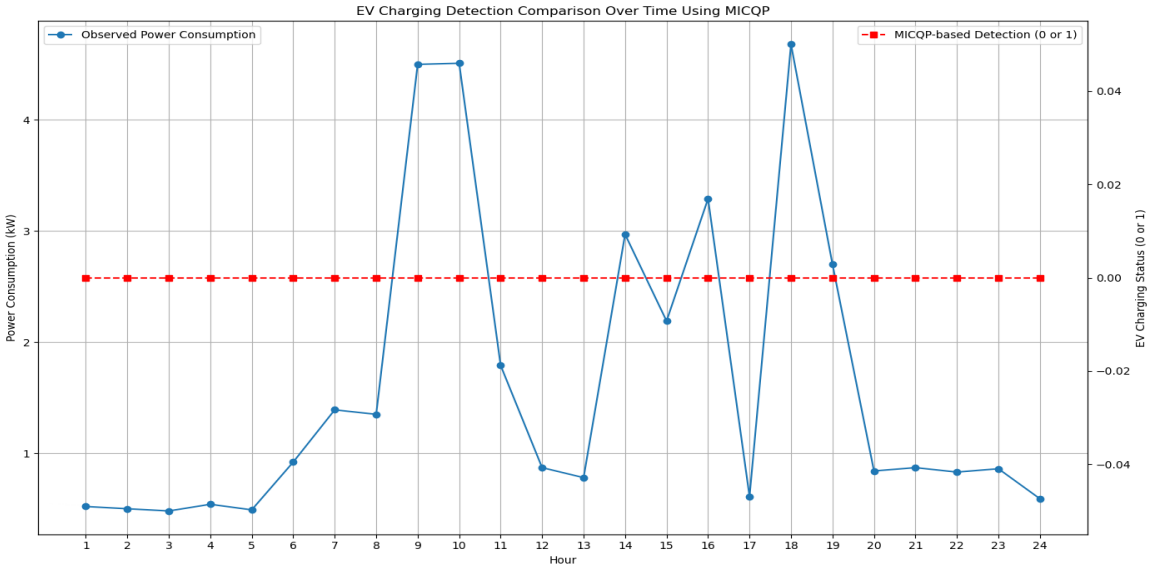


Figure 7: EV Charging Detection Using MICQP

The below figure (Figure 8) presents the probability percentages for EV charging start times. The histogram indicates a higher likelihood of EV charging starting between 7 AM and 9 AM, with additional significant probabilities around 5 PM and 7 PM. These results align with typical residential charging behaviors, where EV owners are likely to charge their vehicles before starting their day or after returning home in the evening.

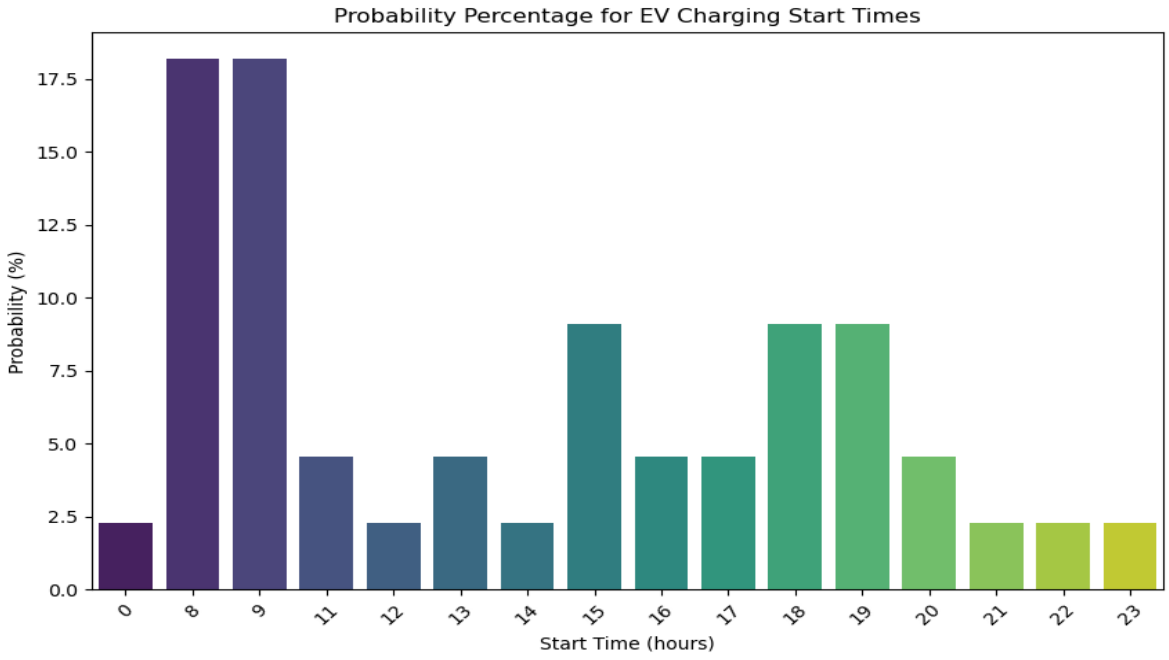


Figure 8: Probability Percentage for EV Charging Start Time

In Figure 9, the probability percentages for the duration of EV charging events. The histogram reveals that most charging sessions last between 1 to 2 hours, with a notable peak at 1 hour. This suggests that most EV owners have short charging sessions, likely due to partial charging or top-up charging during the day.

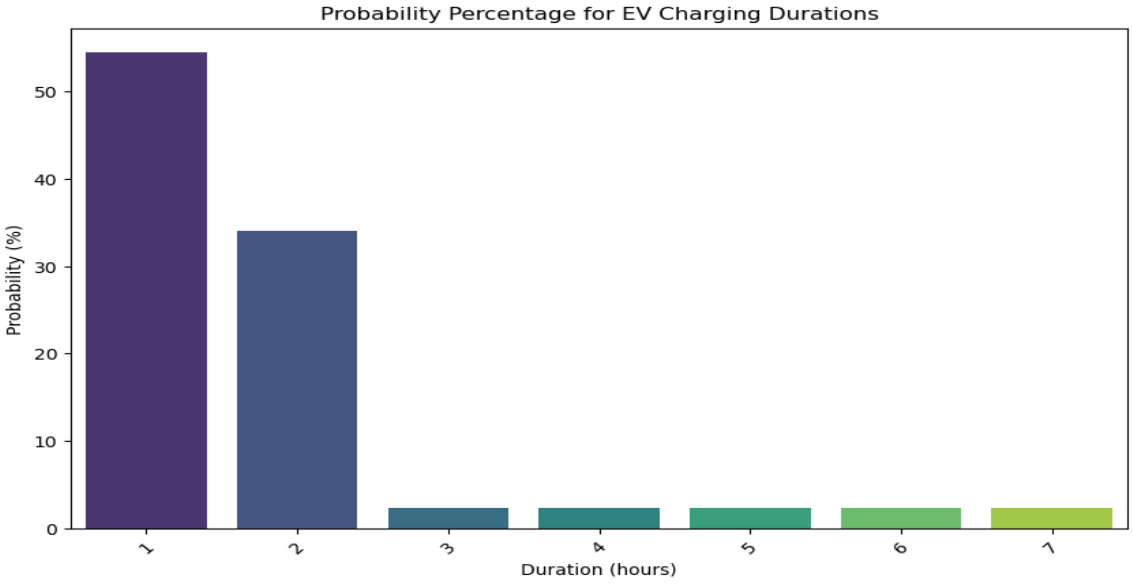


Figure 9: Probability Percentage for EV Charging Start Time

The MICQP model demonstrated effectiveness in identifying EV charging events by capturing power usage patterns associated with charging behaviors. The model's ability to handle both continuous and binary variables allow it to distinguish between regular power consumption and charging events. The results from the probability analyses for charging start times and durations provide insights into common charging behaviors as well. However, it should be noted that the resulting probability pattern doesn't match the usual usage pattern, which typically shows higher consumption between 4 PM and 11 PM, especially during off-peak times. This discrepancy suggests that the solver might have a tolerance level that causes small variations to be overlooked.

5.5. Exponential Moving Average (EMA) and Wavelet Decomposition

In our analysis, we applied both Exponential Moving Average (EMA) and Wavelet Decomposition to identify and understand the EV charging behaviors at various household locations. Using EMA, spikes in power usage were successfully identified on several days, suggesting occasional high power consumption events which may correspond to EV charging.

The EMA technique was employed to smooth the power usage data, allowing us to highlight significant deviations that suggest potential EV charging events. By giving more weight to recent observations, EMAeffectively identifies spikes in power usage. For instance, in the power usage graphs, spikes detected by EMA (indicated by red markers) align with expected high consumption events, likely due to EV charging.

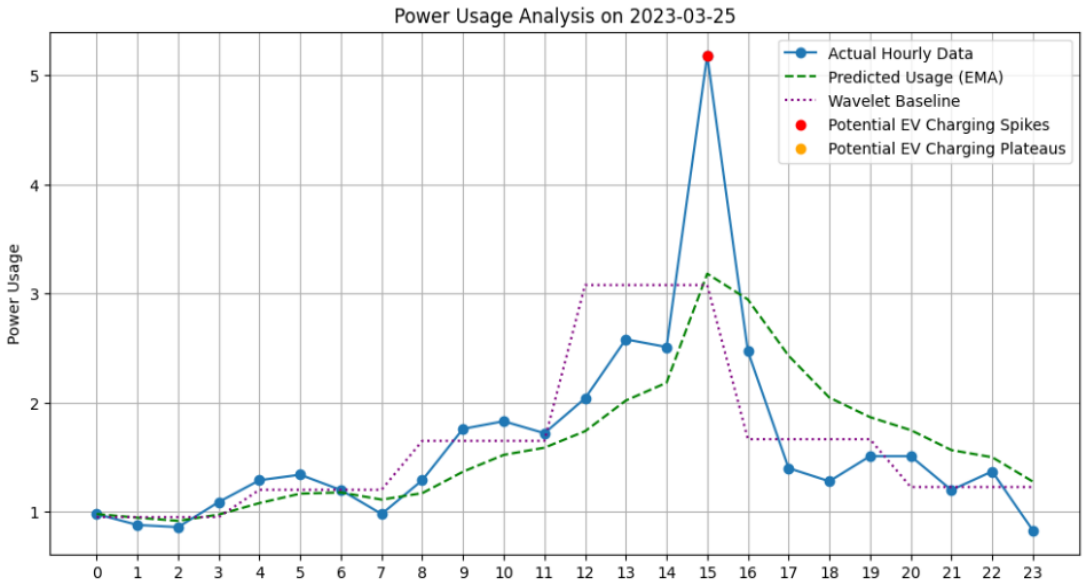


Figure 10: Power Usage Analysis Using EMA

Wavelet Decomposition proved effective in identifying plateaus in power usage, indicative of prolonged periods of high consumption. Some sample result is shown below:

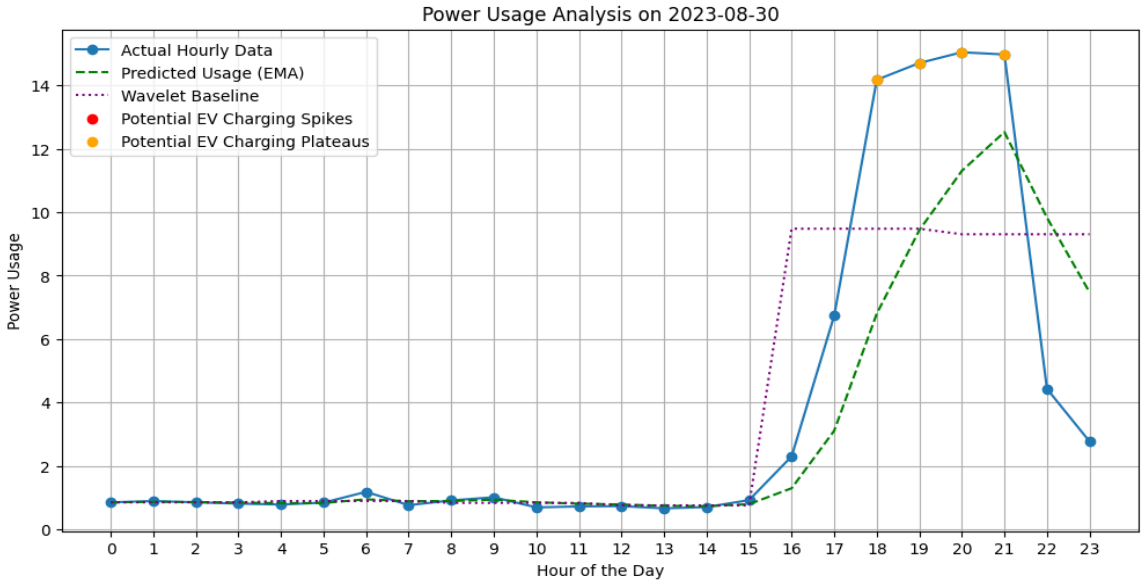


Figure 11: Power Usage Analysis Using Wavelet Decomposition

Wavelet Decomposition, particularly using the Daubechies wavelet ('db1'), enabled us to break down the power usage signal into different components. The low-frequency approximation coefficients were used to reconstruct a smooth baseline, effectively filtering out the noise and highlighting periods of sustained high-power usage (indicated by orange markers). These plateaus are indicative of continuous EV charging sessions.

Below are the sample of the combined result of the whole labelled dataset. As shown in the table, the percentage of charging day of different location (household) is different due to different charging behaviours. A household usually charge their EV within 2 or 4 days.

Table 1: Sample of the combined result of whole labeled

Location	Percentage of Days Charged
23	32.87671233
27	53.42465753
295	12.87671233
1048	53.15068493
1173	6.575342466
1919	93.69863014
2105	49.8630137
2197	28.49315068
2384	52.60273973
2468	47.94520548
2478	71.78082192
2573	75.34246575
2621	38.90410959
2644	34.79452055
2857	42.19178082
2880	56.71232877
2888	51.78082192
2935	21.36986301
2982	49.31506849
3341	38.63013699
3369	71.23287671
3392	80.54794521
3415	56.43835616
3716	29.31506849
3725	66.30136986
3888	8.493150685

And the sample heatmap below shows the charging patterns for different locations. And most of the household will show certain patterns for charging actions. As the results of these two methods aligns with the charging pattern of the typical off-peak charging behavior, we opted to use Wavelet decomposition over MICQP as our final base model.

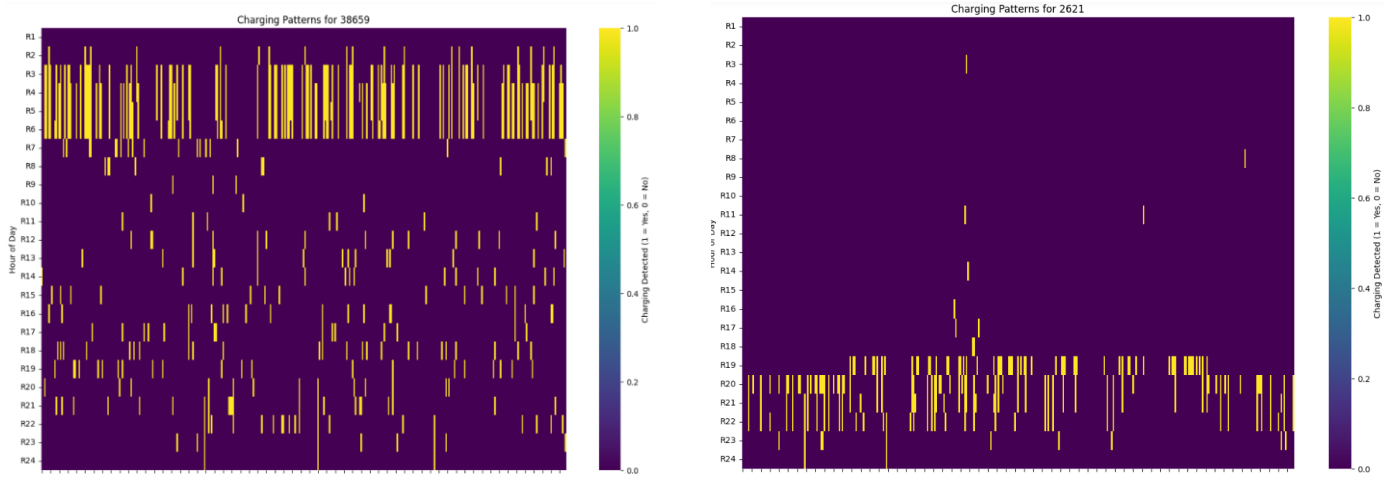


Figure 12: Heatmap showing the charging pattern using Wavelet decomposition

5.6. Time Series Analysis

The fixed rank of average daily power usage of registered EV users is ARIMA (3, 0, 3). The parameters of (3, 0, 3) show the power usage of current timestamp is related to usages and white noises of previous 3 stamps. Also, the regression formula is based on original data without differential operation.

$$d_t = -0.4011 + 0.6127d_{t-1} + 0.7162d_{t-2} - 0.3580d_{t-3} + \varepsilon_t + 0.2844\varepsilon_{t-1} - 0.6627\varepsilon_{t-2} - 0.4093\varepsilon_{t-3},$$
$$\{\varepsilon_t\} \sim^{i.i.d} N(0, 20.0412)$$

The fixed rank of average hourly power usage of registered EV users is SARIMA (2, 0, 1) × (1, 1, 1, 4). The parameters of (2, 0, 1) × (1, 1, 1, 4) show that the hourly power usage records need no one-step differential and 1 s-step differential operation. The length of seasonal periods is 4. After differencing, the data of current timestamp is related to the data of previous 2 stamps and the white noise of previous 1 stamp. Also, the data is related to the data and white noise of previous 1 seasonal period (which is 4 stamps ahead).

$$h_t = 1.5175h_{t-1} - 0.5199h_{t-2} - 0.0578h_{t-4} + \varepsilon_t - 0.9053\varepsilon_{t-1} - 0.9999\varepsilon_{t-4},$$
$$\{\varepsilon_t\} \sim^{i.i.d} N(0,0.0437)$$

The result of ARIMA daily is not that informative or referable. Because the variance of white noises is

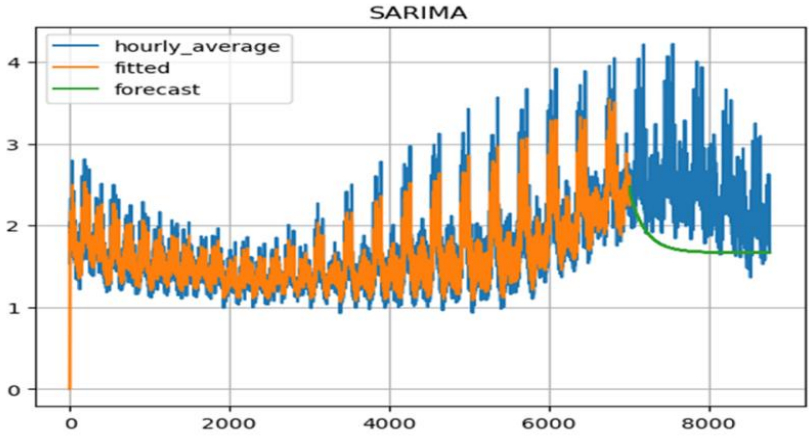


Figure 13: Fitted model on hourly usage and its forecast

quite big in comparison with the coefficients of historical records. Though all parameters are statistically significant, the regression is highly dependent on the randomness part. For more detailed hourly usage data, the result seems more reasonable. However, the SARIMA is still limited in its explanatory and predictive power.

As the plot above shows, the fitted orange line is close to the blue original line. But there is a bigger gap between the green forecasted results and the true value. A more complex regression structure is needed to capture the periodical fluctuations in the data.

To conclude, Time Series Analysis is feasible but not that suitable for residential records. Because the electricity usage for charging EV is relatively stable in one family. No family is likely to buy new EVs then

sell some in our current weekly-scaled data. On the other hand, Time Series Analysis could be more effective in forecasting for public chargers as the paper shows. (K. C. Akshay, 2024) On the plus side, Time Series Analysis gives us a new perspective to understand the original data better. Further on, more complex models with bigger fixed ranks can be explored afterwards to discover patterns in data changes.

5.7. Clustering Algorithm

5.7.1. Initial Clustering Result

The clustering algorithms' performance was assessed using the Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Index. The results are summarized in the table below:

Table 2: Evaluation of clustering algorithms on our dataset

Model Name	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Index
KMeans	0.5149	0.8143	464.5123
Hierarchical	0.5149	0.8143	464.5123
DBSCAN		Only one cluster found	
MeanShift	0.3948	1.0216	64.9731
GMM	0.5101	0.8220	451.8174
Spectral	0.5149	0.8143	464.5123
OPTICS	0.3930	1.2473	293.1408

The greater Silhouette Score (values in [-1, 1]) and Calinski-Harabasz Index (positive values) indicate better clustering effects. The Silhouette Score measures how similar an object is to its own cluster compared to other clusters, with higher values signifying better-defined clusters. The Calinski-Harabasz Index evaluates the ratio of the sum of the between-clusters dispersion and of the within-cluster dispersion, with higher values suggesting that the clusters are dense and well-separated. Conversely, the Davies-Bouldin Index (positive values) is a metric where smaller values indicate better clustering. This index assesses the average similarity ratio of each cluster with the one that is most similar to it, with lower values implying less similarity between clusters and therefore better clustering performance.

Among several algorithms, KMeans and Hierarchical stand out based on these metrics. The KMeans algorithm achieved a Silhouette Score of 0.5149, a Davies-Bouldin Index of 0.8143, and a Calinski-Harabasz Index of 464.5123, indicating well-defined and distinct clusters. Similarly, the Hierarchical clustering algorithm showed identical performance metrics to KMeans, suggesting that it also effectively identified clear and distinct clusters within the data. On the other hand, the DBSCAN algorithm only found one cluster, which indicates that it was not effective for this dataset. MeanShift and OPTICS clustering algorithms showed relatively lower performance, with Silhouette Scores of 0.3948 and 0.3930,

respectively, and higher Davies-Bouldin Index values, suggesting less distinct and more overlapping clusters. The GMM algorithm, while performing slightly better than MeanShift and OPTICS with a Silhouette Score of 0.5101 and a Davies-Bouldin Index of 0.8220, still did not outperform KMeans and Hierarchical clustering. Spectral clustering mirrored the performance of KMeans and Hierarchical, with a Silhouette Score of 0.5149, a Davies-Bouldin Index of 0.8143, and a Calinski-Harabasz Index of 464.5123, reinforcing the effectiveness of these methods in producing high-quality clustering results.

To visualize the clustering result, we took KMeans as an example and combine the usage of 24 dimensions into 3 axes (x: 1:00-8:00, y: 9:00-16:00, and z: 17:00-24:00).

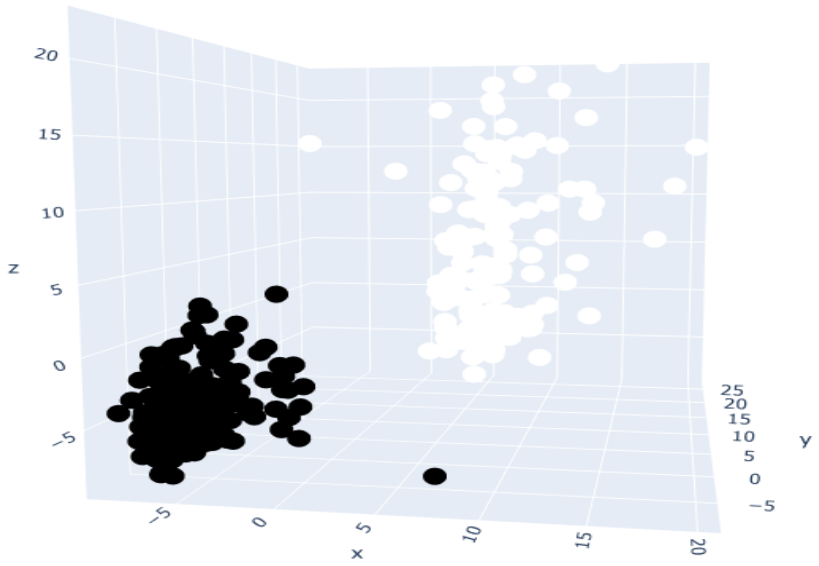


Figure 14: Clusters of usage data for one unique location

The black data points in the plot above correspond to those days with lower power usage (than average level). And the white points correpond to those days with higher power usage. Note that the values of three axes are the sums of every eight hours in order to enable data visualization for a high-dimensional hyperspace. The result is based on an intuitive assumption, that the charging days consume more power than those not charging days. Therefore, we can simply divide the data points into charging days (white) and not charging days (black). Then we can specifically analyze the values of outliers and find out the excessive deviations. At the same time, we explored the clustering results with more clusters shown in Figure 15. The results inspire us to tag different days as different groups (such as higher consumption of the first 8 hours, the medium 8 hours, the last 8 hours, and the higher consumption of the whole day). To conclude, clustering algorithms are powerful in dealing with unlabeled data and labelling it.

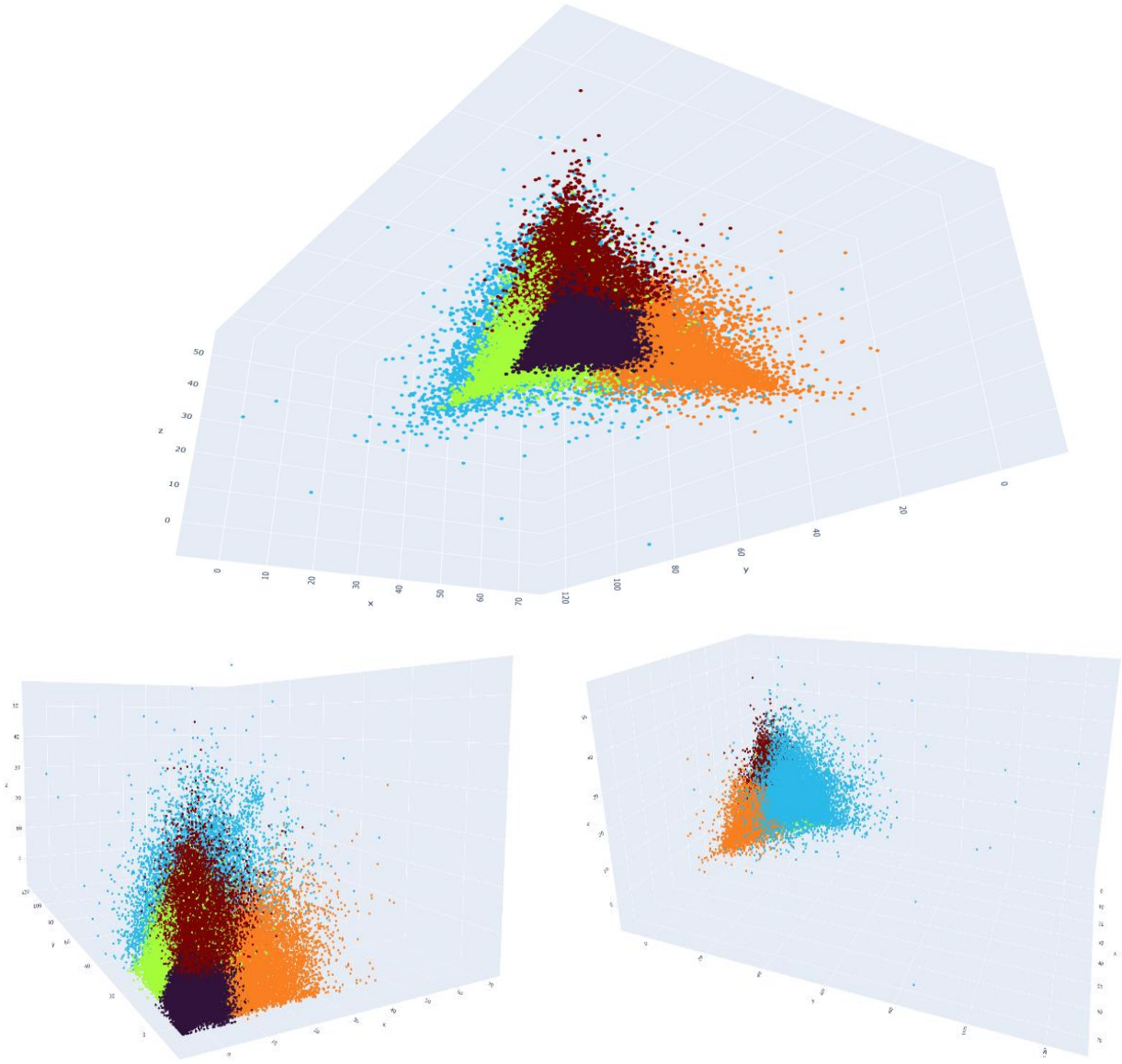


Figure 15: Clusters of all usage

5.7.2. Data Segmentation by Months

To avoid the seasonal variations in power consumption, we segmented the dataset into 12 separate monthly datasets. This step was crucial to ensure that high consumption periods, such as those in summer, did not disproportionately affect the clustering results.

Monthly Segmentation:

- The entire dataset was divided into 12 subsets, each representing one month of the year.
- This segmentation allowed for more accurate analysis of consumption patterns within each

specific period, avoiding the distortion that could arise from combining data across different seasons.

5.7.3. Categorization of Hourly Data

To reflect the varying electricity unit prices and potential usage behaviors, we categorized the 24-hour day into three distinct periods: Off-peak, Mid-peak, and On-peak hours.

Combination Method for Hourly Categorization:

- **Off-peak Hours:** Typically, these hours have the lowest electricity rates and often occur during the night and early morning.
- **Mid-peak Hours:** These hours have moderate electricity rates and occur during transitional periods, such as late morning and early afternoon.
- **On-peak Hours:** These are the hours with the highest electricity rates, usually occurring during the late afternoon and early evening when overall consumption is at its peak.

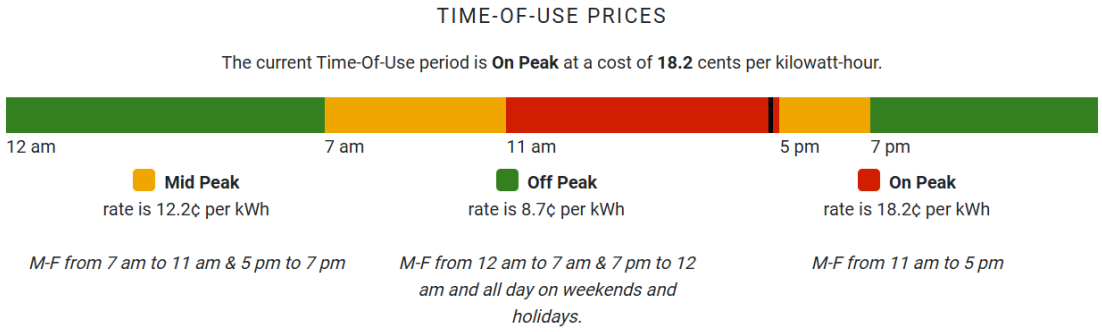


Figure 17: BHI Time-Of- Use Prices

By dividing the day into these three periods, we aimed to capture different usage behaviors corresponding to the varying electricity prices. This method provided a clearer picture of when high power consumption, likely due to EV charging, was occurring.

5.7.4. Clustering Implementation

Clustering Parameters:

- **Cluster Numbers:** For each month's dataset, we set the clustering parameter to 3 or 4 clusters. This number was chosen based on the observed variability in consumption patterns within each period.
- **Algorithm Application:** We applied clustering algorithms to each monthly dataset, analyzing the 24-hour consumption patterns to identify distinct groups of usage behaviors.

Taking January as an example, the clustering algorithm categorized the data into four distinct groups:

- **Cost Sensitive:** High consumption during off-peak hours.
- **Insensitive:** Average consumption across all hours.
- **Thrifty:** Consistently low consumption, approximately 1 kWh.
- **Consuming:** High daily consumption, exceeding 100 kWh.

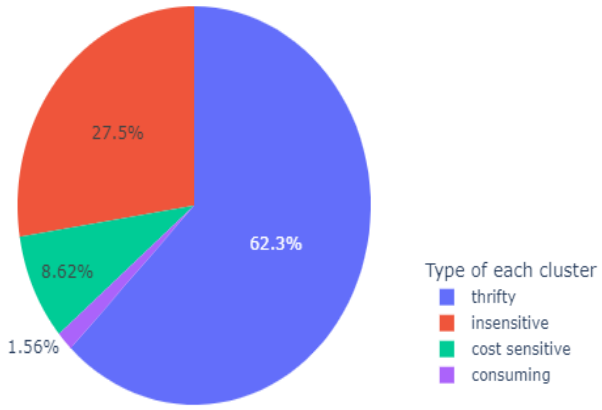


Figure 18: Different clusters for the month of January

The ratios of these clusters were depicted in a pie chart shown in Figure 18, and their data points were shown in a 3D scatter plot (Figure 19), providing a comprehensive visualization of the consumption

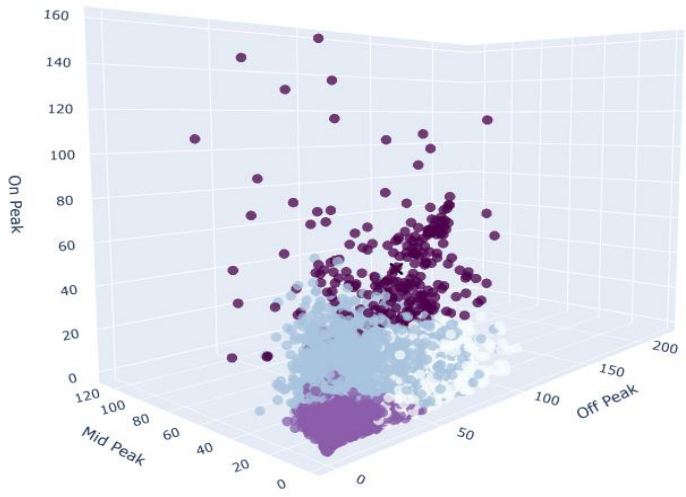
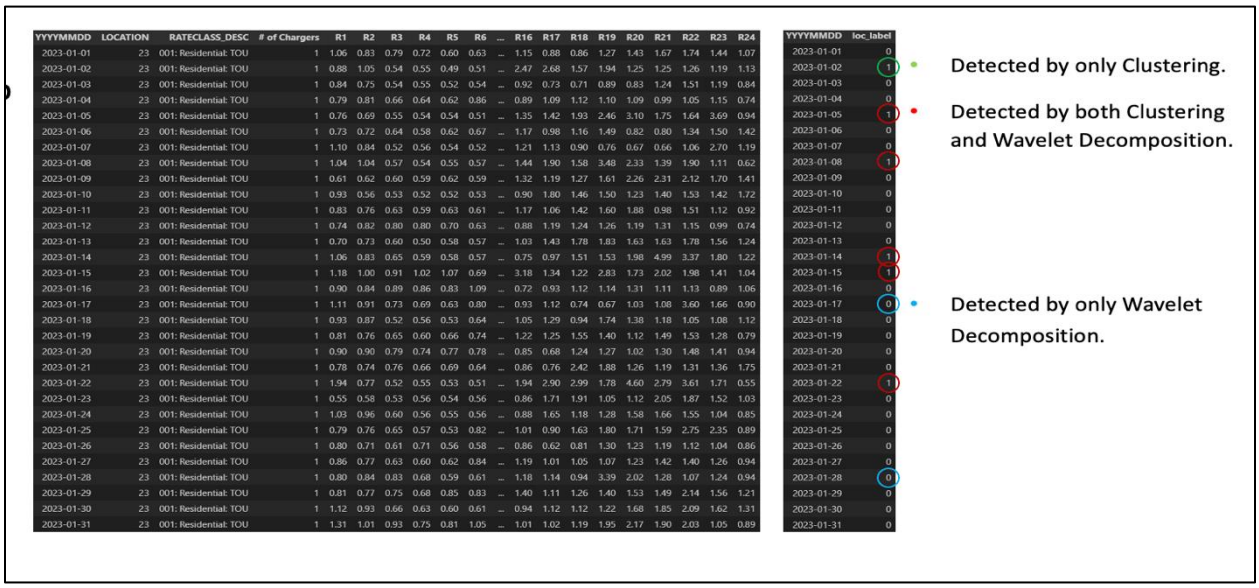


Figure 19: January Data Points

patterns. The scatter plots illustrate the distribution of data points across different months, highlighting the variations in power usage patterns. Each cluster's characteristics, such as "Cost Sensitive," "Insensitive," "Thrifty," and "Consuming," were effectively captured and visualized, offering insights into how power consumption behaviors change over time. These visualizations in the Appendix (Figure 2 and3) support the detailed analysis of the clustering results and provide a clear representation of the seasonal and daily consumption trends.

5.7.5. Second Step Clustering

After the initial clustering, we applied a second step clustering with the hypothesis that charging days would have higher consumption than non-charging days. The results were validated against Wavelet Decomposition and Threshold Policy, enhancing confidence in detecting actual charging days.



YYYYMMDD	LOCATION	RATECLASS_DESC	# of Chargers	R1	R2	R3	R4	R5	R6	...	R16	R17	R18	R19	R20	R21	R22	R23	R24
2023-01-01	23	001: Residential: TOU	1	1.06	0.83	0.79	0.72	0.60	0.63	...	1.15	0.88	0.86	1.27	1.43	1.67	1.74	1.44	1.07
2023-01-02	23	001: Residential: TOU	1	0.88	1.05	0.54	0.55	0.49	0.51	...	2.47	2.68	1.57	1.94	1.25	1.25	1.26	1.19	1.13
2023-01-03	23	001: Residential: TOU	1	0.84	0.75	0.54	0.55	0.52	0.54	...	0.92	0.73	0.71	0.89	0.83	1.24	1.51	1.19	0.84
2023-01-04	23	001: Residential: TOU	1	0.79	0.81	0.66	0.64	0.62	0.86	...	0.89	1.09	1.12	1.10	1.09	0.99	1.05	1.15	0.74
2023-01-05	23	001: Residential: TOU	1	0.76	0.69	0.55	0.54	0.54	0.51	...	1.35	1.42	1.93	2.46	3.10	1.75	1.64	3.69	0.94
2023-01-06	23	001: Residential: TOU	1	0.73	0.72	0.64	0.58	0.62	0.67	...	1.17	0.98	1.16	1.49	0.82	0.80	1.34	1.50	1.42
2023-01-07	23	001: Residential: TOU	1	1.10	0.84	0.52	0.56	0.54	0.52	...	1.21	1.13	0.90	0.76	0.67	0.66	1.06	2.70	1.19
2023-01-08	23	001: Residential: TOU	1	1.04	1.04	0.57	0.54	0.55	0.57	...	1.44	1.90	1.58	3.48	2.33	1.39	1.90	1.11	0.62
2023-01-09	23	001: Residential: TOU	1	0.61	0.62	0.60	0.59	0.62	0.59	...	1.32	1.19	1.27	1.61	2.26	2.31	2.12	1.70	1.41
2023-01-10	23	001: Residential: TOU	1	0.93	0.56	0.53	0.52	0.52	0.53	...	0.90	1.80	1.46	1.50	1.23	1.40	1.53	1.42	1.72
2023-01-11	23	001: Residential: TOU	1	0.83	0.76	0.63	0.59	0.63	0.61	...	1.17	1.06	1.42	1.60	1.88	0.98	1.51	1.12	0.92
2023-01-12	23	001: Residential: TOU	1	0.74	0.82	0.80	0.80	0.70	0.63	...	0.88	1.19	1.24	1.26	1.19	1.31	1.15	0.99	0.74
2023-01-13	23	001: Residential: TOU	1	0.70	0.73	0.60	0.50	0.58	0.57	...	1.03	1.43	1.78	1.83	1.63	1.63	1.78	1.56	1.24
2023-01-14	23	001: Residential: TOU	1	1.06	0.83	0.65	0.59	0.58	0.57	...	0.75	0.97	1.51	1.53	1.98	4.99	3.37	1.80	1.22
2023-01-15	23	001: Residential: TOU	1	1.18	1.00	0.91	1.02	1.07	0.69	...	3.18	1.34	1.22	2.83	1.73	2.02	1.98	1.41	1.04
2023-01-16	23	001: Residential: TOU	1	0.90	0.84	0.89	0.86	0.83	1.09	...	0.72	0.93	1.12	1.14	1.31	1.11	1.13	0.89	1.06
2023-01-17	23	001: Residential: TOU	1	1.11	0.91	0.73	0.69	0.63	0.80	...	0.93	1.12	0.74	0.67	1.03	1.08	3.60	1.66	0.90
2023-01-18	23	001: Residential: TOU	1	0.93	0.87	0.52	0.56	0.53	0.64	...	1.05	1.29	0.94	1.74	1.38	1.18	1.05	1.08	1.12
2023-01-19	23	001: Residential: TOU	1	0.81	0.76	0.65	0.60	0.66	0.74	...	1.22	1.25	1.55	1.40	1.12	1.49	1.53	1.28	0.79
2023-01-20	23	001: Residential: TOU	1	0.90	0.90	0.79	0.74	0.77	0.78	...	0.85	0.68	1.24	1.27	1.02	1.30	1.48	1.41	0.94
2023-01-21	23	001: Residential: TOU	1	0.78	0.74	0.76	0.66	0.69	0.64	...	0.86	0.76	2.42	1.88	1.26	1.19	1.31	1.36	1.75
2023-01-22	23	001: Residential: TOU	1	1.94	0.77	0.52	0.55	0.53	0.51	...	1.94	2.90	2.99	1.78	4.60	2.79	3.61	1.71	0.55
2023-01-23	23	001: Residential: TOU	1	0.55	0.58	0.53	0.56	0.54	0.56	...	0.86	1.71	1.91	1.05	1.12	2.05	1.87	1.52	1.03
2023-01-24	23	001: Residential: TOU	1	1.03	0.96	0.60	0.56	0.55	0.56	...	0.88	1.65	1.18	1.28	1.58	1.66	1.55	1.04	0.85
2023-01-25	23	001: Residential: TOU	1	0.79	0.76	0.65	0.57	0.53	0.82	...	1.01	0.90	1.63	1.80	1.71	1.59	2.75	2.35	0.89
2023-01-26	23	001: Residential: TOU	1	0.80	0.71	0.61	0.71	0.56	0.58	...	0.86	0.62	0.81	1.30	1.23	1.19	1.12	1.04	0.86
2023-01-27	23	001: Residential: TOU	1	0.86	0.77	0.63	0.60	0.62	0.84	...	1.19	1.01	1.05	1.07	1.23	1.42	1.40	1.26	0.94
2023-01-28	23	001: Residential: TOU	1	0.80	0.84	0.83	0.68	0.59	0.61	...	1.18	1.14	0.94	3.39	2.02	1.28	1.07	1.24	0.94
2023-01-29	23	001: Residential: TOU	1	0.81	0.77	0.75	0.68	0.85	0.83	...	1.40	1.11	1.26	1.40	1.53	1.49	2.14	1.56	1.21
2023-01-30	23	001: Residential: TOU	1	1.12	0.93	0.66	0.63	0.60	0.61	...	0.94	1.12	1.12	1.22	1.68	1.85	2.09	1.62	1.31
2023-01-31	23	001: Residential: TOU	1	1.31	1.01	0.93	0.75	0.81	1.05	...	1.01	1.02	1.19	1.95	2.17	1.90	2.03	1.05	0.89

YYYYMMDD	loc_label
2023-01-01	0
2023-01-02	1
2023-01-03	0
2023-01-04	0
2023-01-05	1
2023-01-06	0
2023-01-07	0
2023-01-08	1
2023-01-09	0
2023-01-10	0
2023-01-11	0
2023-01-12	0
2023-01-13	0
2023-01-14	1
2023-01-15	1
2023-01-16	0
2023-01-17	0
2023-01-18	0
2023-01-19	0
2023-01-20	0
2023-01-21	0
2023-01-22	1
2023-01-23	0
2023-01-24	0
2023-01-25	0
2023-01-26	0
2023-01-27	0
2023-01-28	0
2023-01-29	0
2023-01-30	0
2023-01-31	0

Figure 20: Second Step Clustering to detect actual charging days

This comprehensive approach ensured that the clustering results were robust, reflective of true usage patterns, and useful for identifying EV charging behaviors.

6. Future Direction

The challenges present a coincident correspondence. Initially, we had comprehensive residential records, with some indicating at least one registered EV. Our objective was to identify non-registered EV users within these records. This task involves pattern recognition with incomplete labels regarding EV ownership. Furthermore, it is necessary to detect all charging days and hours, despite incomplete labels indicating whether charging occurred on a given day. To address this "downscaled" problem, deep learning models can be employed. These models will help identify non-registered EV users and, by including transformer load data, forecast future capacity needs, thereby guiding infrastructure enhancements to support increased EV charger connections.

In pursuit of model refinement and training, developing, and training advanced deep learning models, such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, can significantly improve the identification of EV charging patterns and non-registered EV users. Extensive hyperparameter tuning should be conducted to optimize the performance of these models, ensuring the best possible accuracy and efficiency. Additionally, exploring advanced analytical techniques, including the application of reinforcement learning, can dynamically adapt the models based on real-time data inputs and evolving usage patterns. Transfer learning can also be utilized to leverage pre-trained models on similar datasets, enhancing the performance of current models with minimal additional training. By integrating these methodologies, the precision and reliability of EV charging load forecasts can be

enhanced, ultimately supporting the efficient and sustainable management of electrical load distribution for Burlington Hydro Inc.

7. Conclusion

This study has demonstrated the effectiveness of advanced data analysis and machine learning techniques in forecasting electric vehicle (EV) charging loads for Burlington Hydro Inc. By integrating Random Forest, Logistic Regression, semi-supervised learning, and neural networks, we achieved accurate detection and classification of EV charging events from extensive power consumption data. Additionally, the application of Mixed-Integer Quadratic Programming (MIQP) and Isolation Forest models provided robust methods for optimizing EV charging load forecasts and identifying anomalies.

The Exponential Moving Average (EMA) successfully identified spikes in power usage, correlating with EV charging events. Wavelet Decomposition effectively highlighted plateaus in power usage, indicating prolonged charging periods. These methods were essential in pinpointing specific EV charging patterns and enhancing detection accuracy.

Our detailed analysis using clustering algorithms, including KMeans and Hierarchical clustering, further categorized days based on power consumption patterns. This approach was instrumental in understanding and differentiating EV charging behaviors, especially when combined with the cross-validation from Wavelet Decomposition and Threshold Policy methods.

While Time Series Analysis offered insights into power usage trends, it showed limitations in predicting residential EV charging data but could be useful for public charging station forecasts. The analysis revealed that charging patterns typically align with off-peak hours, supporting the decision to prioritize Wavelet Decomposition over Mixed-Integer Convex Quadratic Programming (MICQP) as our base model for final analysis.

This comprehensive approach not only aids Burlington Hydro Inc. in efficiently managing electrical load distribution but also supports environmental sustainability by promoting the transition to a low-carbon economy. The methodologies and findings from this study lay a solid foundation for future research and development of intelligent EV charging infrastructure, ensuring grid reliability, optimizing resource allocation, and fostering sustainable urban growth.

In conclusion, our study demonstrates the potential of using advanced analytical and machine learning techniques to improve EV charging load forecasts, detect anomalies, and categorize usage patterns. These findings provide Burlington Hydro Inc. with valuable insights and tools to manage and optimize their electrical grid in response to increasing EV adoption, contributing to a more sustainable and reliable energy future.

8. References

A.K. Jain, M. M.-3. (1999). Data clustering: A review. *ACM Journals , ACM computing surveys*, 264-323.

Davis, P. J. (2016). *Introduction to Time Series and Forecasting.* Springer.

Ester, M. K.-P. (1996).). A density-based algorithm for discovering clusters in large spatial databases with noise. 226-231.

Feng Li, É. C.-L. (2024). Inferring electric vehicle charging patterns from smart meter data for impact studies. *Les Cahiers du GERAD, (G–2024–02). GERAD, HEC Montréal*(ISSN: 0711-2440), https://www.gerad.ca/en/papers/G-2024-02.

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *IEEE*, 90-95.

Hyndman, R. &. (2013). *Forecasting: Principles and Practice.* OText.

IBM. (2021, November 25). *MIQP: Mixed integer programs with quadratic terms in the objective function. Retrieved*. (IBM) Retrieved from https://www.ibm.com/docs/en/icos/20.1.0?topic=smippqt-miqp-mixed-integer-programs-quadratic-terms-in-objective-function

IBM. (2023, December 12). *What is semi-supervised learning?* (IBM) Retrieved from https://www.ibm.com/topics/semi-supervised-learning

Jiawei Han, M. K. (2011). *Data Mining: Concepts and Techniques.* Elsevier .

K. C. Akshay, G. H. (2024, March 18). Power consumption prediction for electric vehicle charging stations and forecasting income. *Nature Scientific Report, Scientific Reports, 14(1)*(6497–6497), https://doi.org/10.1038/s41598-024-56507-2 . Retrieved from www.nature.com/scientificreports: https://www.nature.com/articles/s41598-024-56507-2

Mallat, S. G. (1999). *A Wavelet Tour of Signal Processing.* Academic Press.

Wolsey, G. N. (1999). *Integer and Combinatorial Optimization.* Wiley-Interscience.

9. Appendix

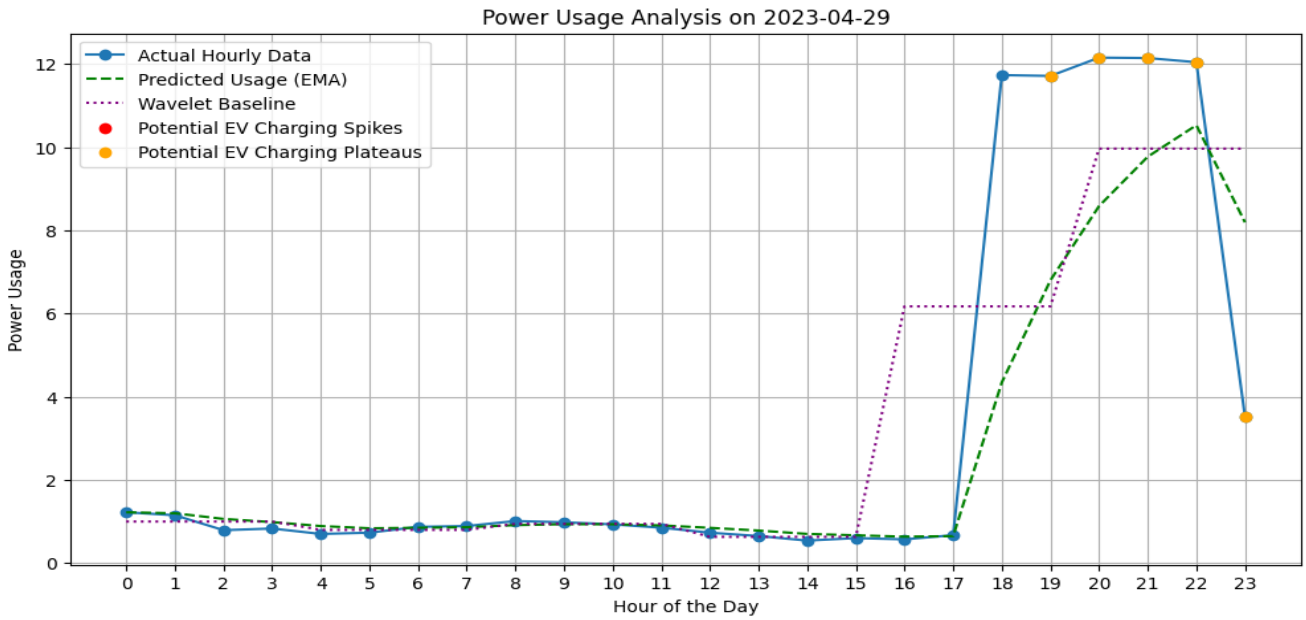


Figure 1: Power usage analysis using Wavelet Decomposition

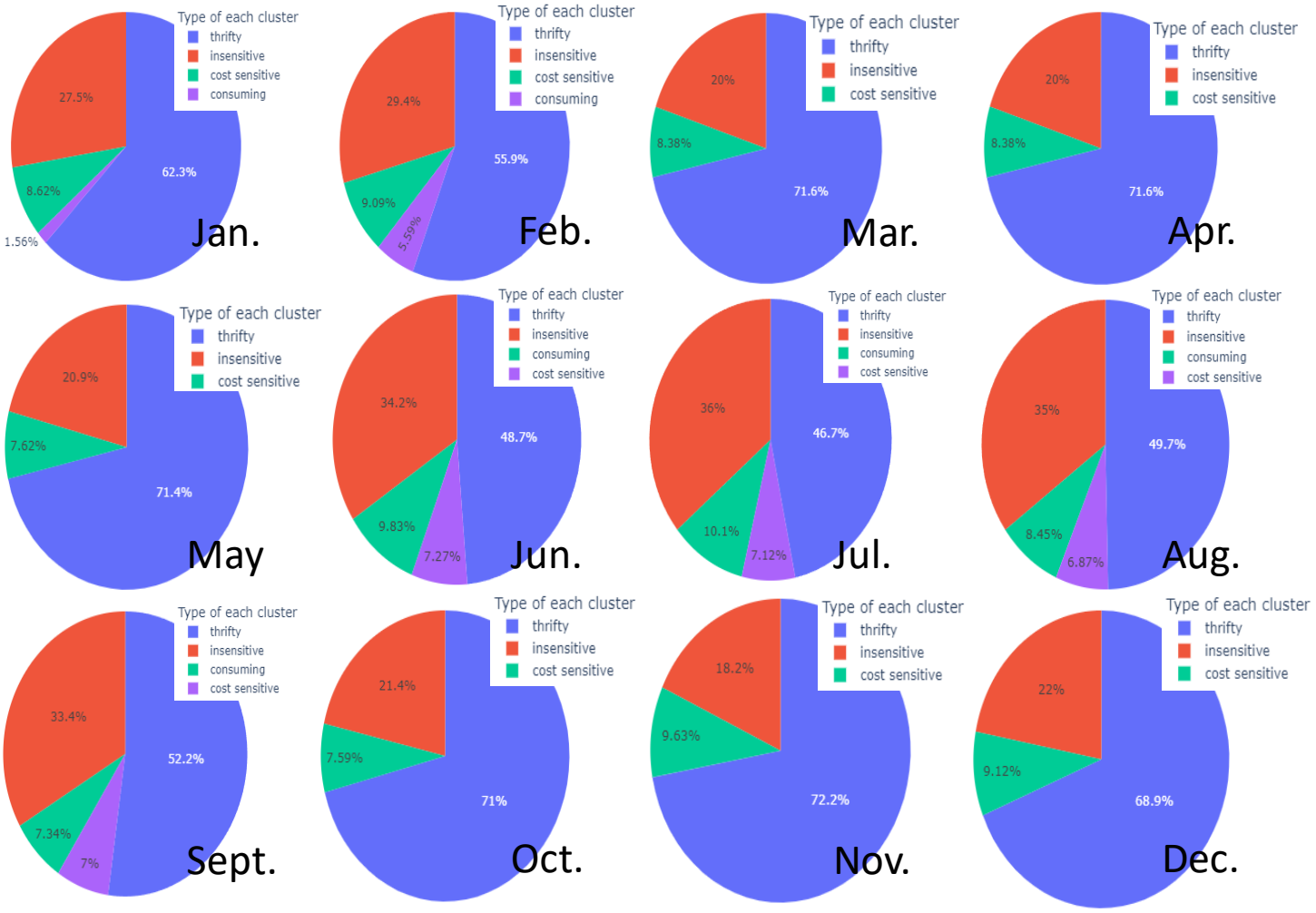


Figure 2: Different clusters for each month of 2023

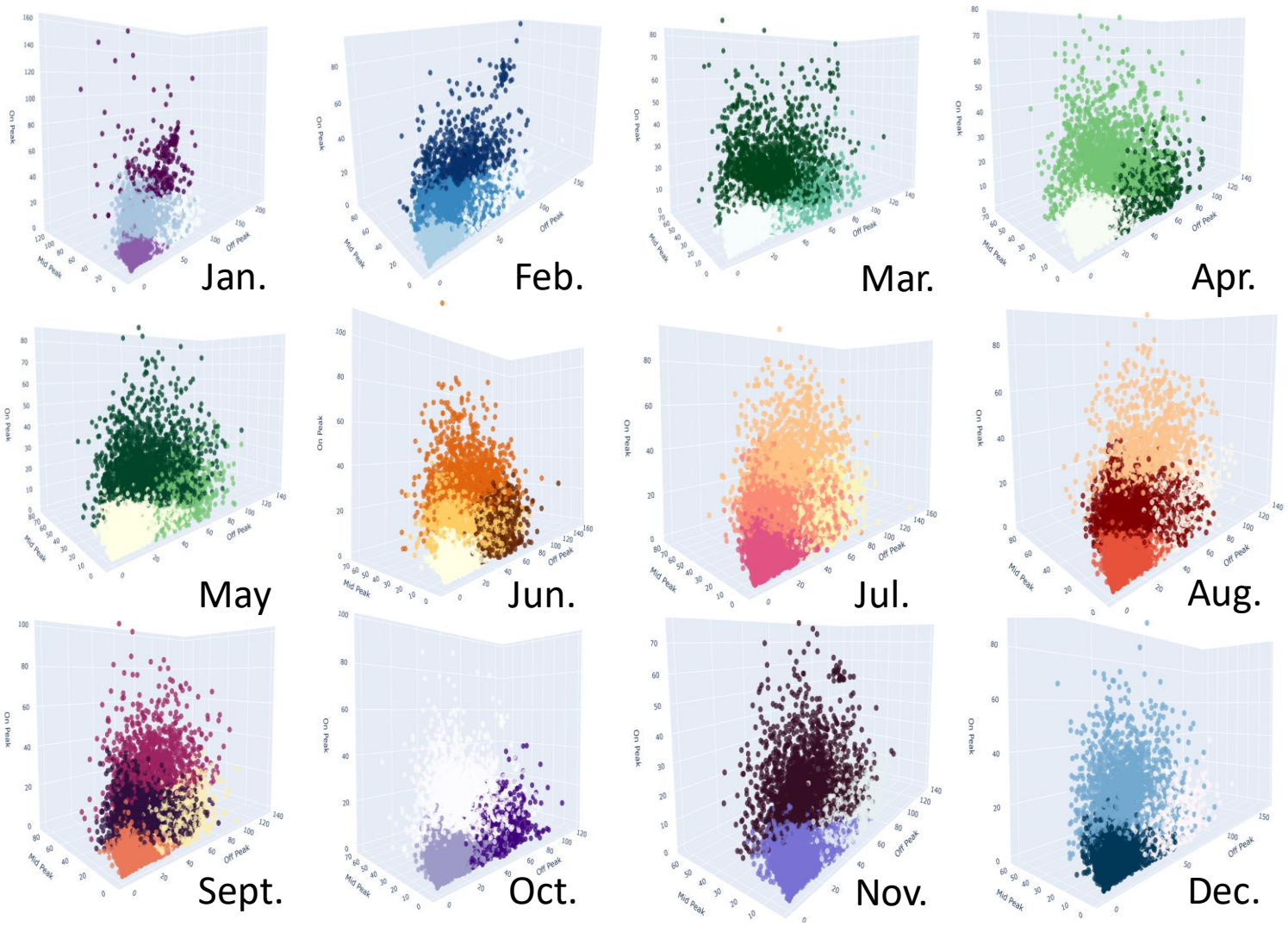


Figure 3: Datapoint distribution for each month of 2023