香港中文大學(深圳)
The Chinese University of Hong Kong, Shenzhen

# Project Report for CSC4008

*Explore Pecan Street Electricity Usage Based on Different Clustering Methods and Entropy*

Xie Huiyu 118010350

Hong Haoyang 118010096

School of Data Science

May 26, 2021

# Contents

# 1 Introduction

This project is aimed at exploring insightful knowledge given **Pecan Street** data. Since a lot of different papers about predicting the electricity consumption have been published, the project was planned for detecting uncovered facts about electricity consumption pattern rather than improving prediction or clustering methods for a better performance.

In this project, multiple clustering methods were compared based on several metrics. The goal is to first detect outliers from the data and extract information from them. Then cluster the electricity usage patterns among each day of one week to see whether there exist distinguished patterns in correlation with weekdays or weekends. Through the clustering, interesting similarities and differences between weekdays and weekends have been discovered.

During the discovery process, an interesting fact is found that the higher the average consumption is, the higher the fluctuation level of the consumption is. To further explore this relationship, a concept called **Entropy** has been introduced and .

# 2 Literature review

Researchers from around the world have used Pecan Street data to publish more than 150 peer-reviewed papers on various topics. As two significant components of data mining, prediction and clustering are two well-worked fields. Therefore, our literature review focused on these two parts. We discovered certain gaps during the review and targeted to address some of them.

## 2.1 Prediction

This section presents some of the work related to prediction utilizing the Pecan Street data. As an essential part of data mining, predicting future consumption is a fruitful field and a myriad of methods have been applied. For example, in [1], three different machine learning models, random forest (RF), multi-layer perception (MLP) and Support Vector Regression (SVR) were compared to a baseline model, persistent forecasting model (PSS), which simply predicts that usage in period t will be equal to usage in period t-1. The author claimed that even computationally-constrained machine learning models perform well than the baseline.

What's more, a hybrid method linearly combining seasonal auto-regressive integrated moving average (SARIMA) model and long-short term memory (LSTM) has been addressed in [2]. SARIMA is a time series model takes seasonal factors into consideration while LSTM can efficiently circumvent back-propagation problems. The trained hybrid method marginally extends the prediction accuracy.

In addition, the authors in [3] addressed another time series model, STL, to put forward the forecasting of monthly power consumption. Combining STL with trend decomposition method X12, the applicability and accuracy are verified through experiments. Furthermore, advanced machine learning methods such as neural network (NN) and deep learning are utilized to achieve better prediction performance.

In [4], the authors proposed a neural network based optimization approach to forecast the energy usage of consumer. Through the methods, both short-term and long-term load

forecasting are addressed. For daily and monthly prediction task, the NNGA obtains high accuracy in the prediction while NNPSO outperforms other method in the field of long-term prediction.

Deep extreme learning machine (DELM) also achieves better results for short-term (one-week) and long-term (one-month) hourly energy consumption prediction than artificial neural network (ANN) and ANFIS in [5]. For training DELM model, different numbers of neurons in the hidden layer and different activation functions have been tuned. And for comparison among different models, different statistical methods such as MAE, MAPE and RMSE are applied as criterions.

## 2.2 Clustering

This section will show the latest research utilizing clustering methods on the Pecan street data. Some research conducted clustering based on the traditional **k-means** method, which was commonly used in data mining. Via using electricity data from 103 homes in Austin, TX [6], homes with similar hourly electricity use patterns were clustered into groups through k-means method. This analysis found that Austin homes fall into one of two seasonal groups with some homes using more expensive electricity.

In [7], all the data from 101 homes in Austin, TX. were used to determine average seasonal profiles in four seasons respectively. Through k-means clustering, all houses were categorized into one cluster. It found that a low percent of the consumers did not change their patterns of electricity usage while the majority of the users changed their electricity usage pattern along with the changing seasons.

Based on daily household electricity use data at an hourly resolution over the year of 2016 across 340 households, [8] utilized unsupervised K-means clustering to cluster households based on 3 evaluation metrics, and further clustered typical use cases of those households at a small $K$. The results of this study are more useful for price modeling and demand response programs than for finding relationships between these demographic factors and energy use.

Some other researches used integrated method to conduct clustering, which overcame the problems appear in traditional clustering methods. [9] proposed an integrated approach which includes two parts: a new method incorporating with community detection (CICD) to improve the LC clustering performance, and a best-cluster-number-determined approach to balance the trade-off between variances within a cluster and further calculate the cluster numbers. This integrated method solved the problem before, ignoring the inherent relationship among different LCs (Load Curves) and the significant volatility and uncertainty of LCs.

Some other research applied more effective methods. In [10], an affinity propagation (AP) algorithm is used to cluster customer load data and generate typical load profiles (TLP) for clusters. AP is an advanced algorithm and has no need for a predefined cluster numbers, which makes the procedure of clustering more effective. Clustering results are compared with some traditional methods such as k means, k-medoid and spectral clustering.

[11] divided k-means clustering into two stage, which improved the clustering quality. It introduced a two-stage method that more accurately captures load shape temporal patterns and peak demands. The proposed approach utilized overpopulation and merging as a method to improve cluster quality, measured by correlation between cluster centroids and individual members.

What's more, other researches achieve improvements in different ways. [12] proposes a flexible DR scheme in smart grid with clustering of residential customers and comprehensively considering the aforementioned factors. New features are extracted from historical data to depict customers' characteristics and clustering methods are applied to explore their electricity consumption habits. Then the clustering information is utilized to help schedule the residential appliances more flexibly and effectively.

In [13], the authors proposed a motif-based association rule mining and clustering technique for determining the energy usage patterns for smart meter data. Further, clustering on the motifs is performed, which gives different consumption behaviors of consumers on different days which can help distribution network operator (DNO) for electricity network modeling and management.

# 3 DataSet Description

## 3.1 Overall Illustration

The data is extracted from Pecan Street data. It measures circuit-level electricity use and generation every minute of every day from nearly 1,000 volunteer homes across their research network [14]. Data used in our project consists of three attributes, 'localminute', 'dataid' and 'use'.
Precisely, the attributes are defined as follows [15]:

- **localminute** stores the time of every electricity usage record.
- **dataid** represents the unique identifier for the home resident pair and if the resident moves, the data collected from the house is associated with a new data ID.
- **use** records the whole home electricity usage within each minute.

The distribution of usage summation for each user is listed as follows:

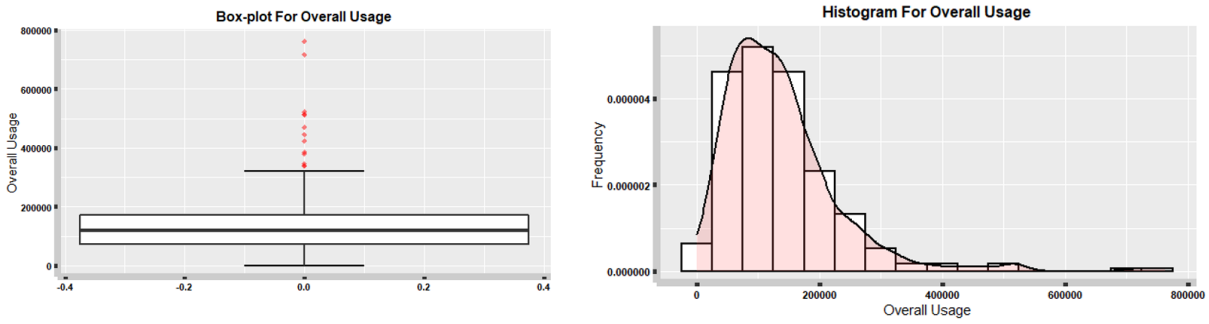



Figure 1: Boxplot and Histogram for Overall Usage

According to the figure, it is shown that there are several users with total 0 electricity consumption. In addition, there is an extremely huge consumption pattern whose usage summation can reach to $800,000$. Most of the users have their total consumption between $5,000$ and $400,000$.

## 3.2 Data Preprocessing

### Outlier Detection

**1. All zero data**. Four users with all 0 records, whose user ids are 2461, 2510, 4830, 7017.

**2. Overly large data**. Cluster the data with average daily electricity consumption and set the number of clusters as 4, the following results were shown.

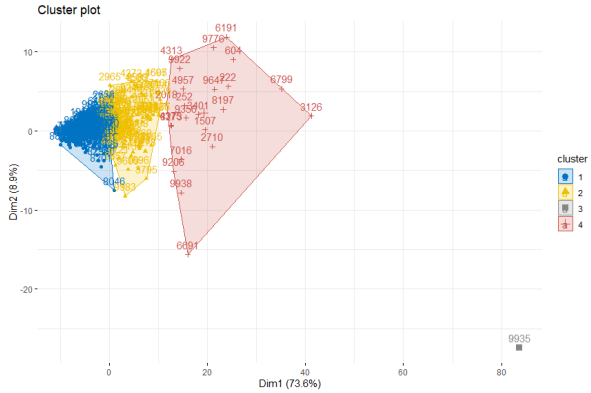



Figure 2: K-means Clustering

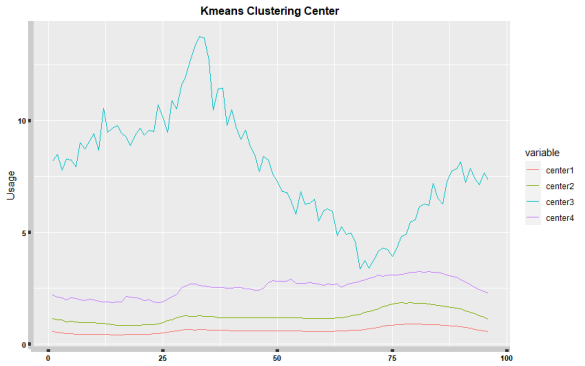
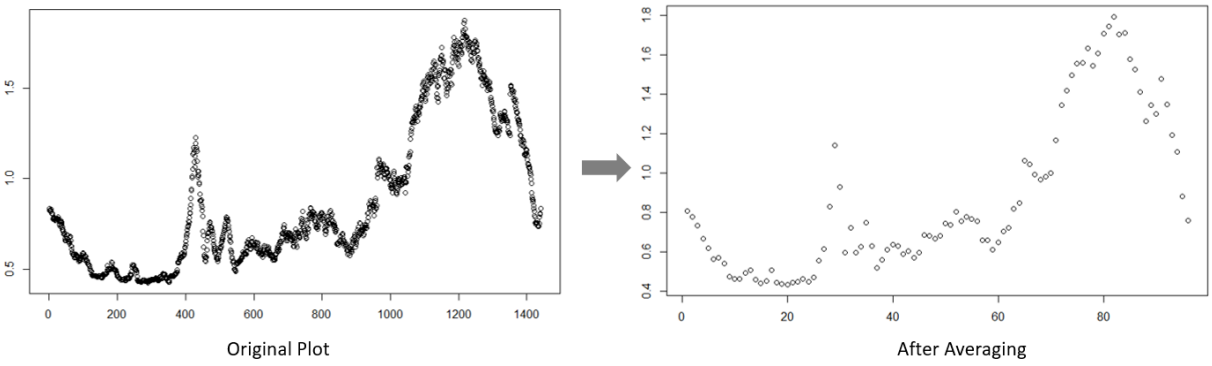



Figure 3: K-means Center Plot

It is shown that there is only one user in cluster 3, which is the highest plot in Fig.3. It obtains extremely large electricity consumption.

### Data Preprocessing

- Eliminate the user with all 0 data and the extremely large electricity consumption.
- Take the granularity as 15-min, i.e. take the average of electricity consumption within each 15 minutes. In this case, not only the basic information will not be erased but save the computation complexity. Select one simple to illustrate.




Then for each user, one vector with length $\frac{24\times60}{15} = 96$ is obtained and will be utilized as the element for different clustering. The normalization part was skipped because the actual electricity consumption can reflect more usage pattern than the data between 0 and 1.

# 4 Different Clustering Methods

This project applied six different (K-means, K-medoids, Soft-Kmeans, Hierarchical Clustering, Hierarchical K-means Clustering, Model Based) clustering methods to the dataset and obtained different results. Compare the performance based on average Silhouette and Dunn Index criterion.

## 4.1 Six Clustering Methods

### K-means

K-means has been widely used as a fundamental clustering method. First to determine the best clustering numbers with respect to gap statistic.

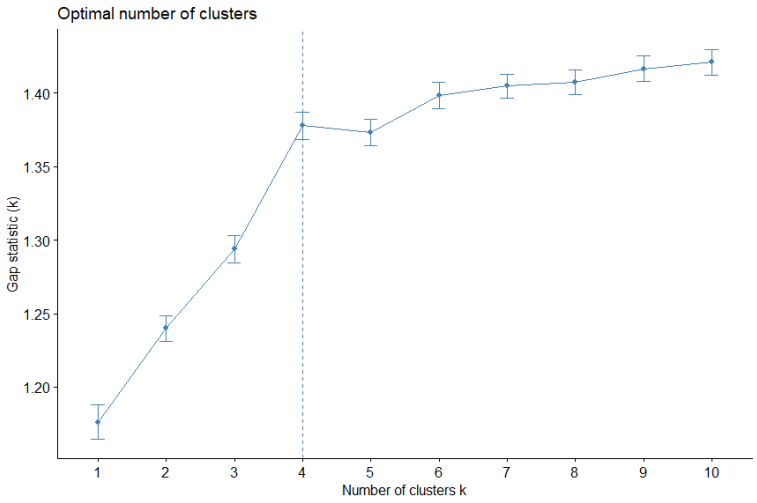



Figure 4: Gap statistic with respect to Number of Clustering

According to the figure, $k = 4$ is the best option. Then conduct k-means with $k = 4$, the clustering result can be shown as follows.

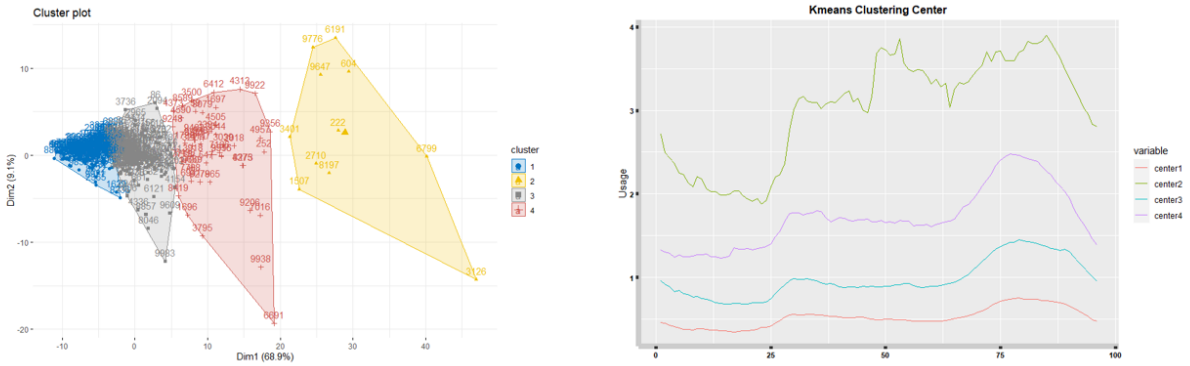



Figure 5: Results for K-means

The first plot is the visualization utilizing the most important two vectors from PCA while the second plot is the center plot for each clustering.

## K-medoids

K-medoids problem is a clustering problem similar to k-means. In contrast to the K-means algorithm, K-medoids chooses actual data points as centers (medoids or exemplars), and thereby allows for greater interpretability of the cluster centers than in K-means, where the center of a cluster is not necessarily one of the input data points (it is the average between the points in the cluster). Furthermore, K-medoids can be used with arbitrary dissimilarity measures, whereas k-means generally requires Euclidean distance for efficient solutions. In this project, the number of clustering is also set as 4, the result is as follows.

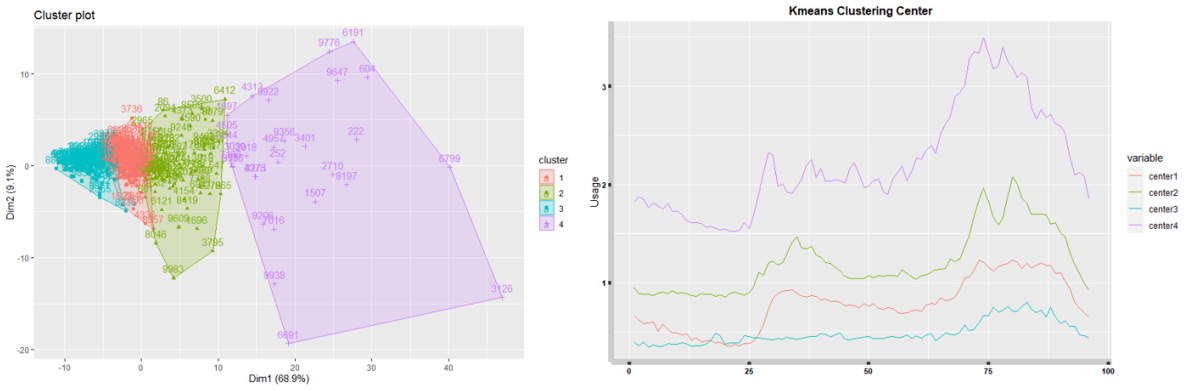



Figure 6: Results for K-medoids

## Soft K-means

Soft k-means is a form of clustering in which each data point can belong to more than one cluster.

Clustering or cluster analysis involves assigning data points to clusters such that items in the same cluster are as similar as possible, while items belonging to different clusters are as dissimilar as possible.

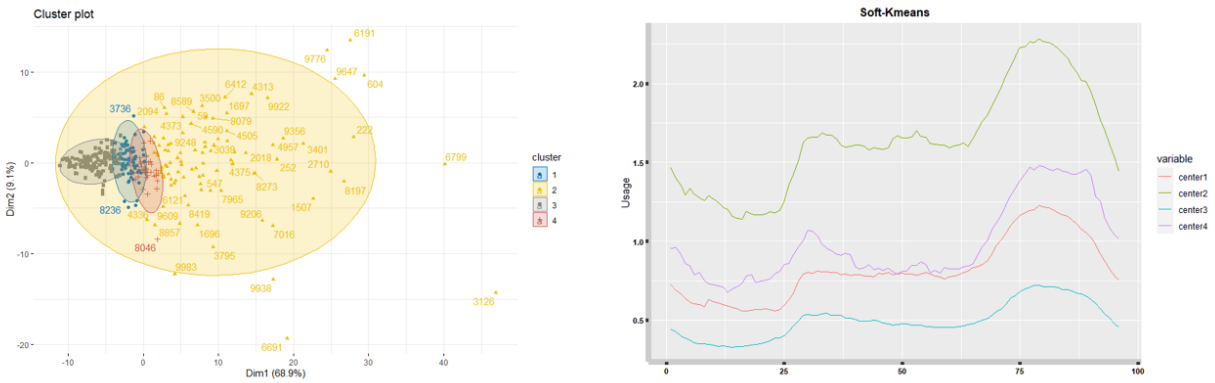



Figure 7: Results for Soft K-means

### Hierarchical Clustering

In data mining and statistics, hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. The key operation in hierarchical agglomerative clustering is to repeatedly combine the two nearest clusters into a larger cluster. The result can be shown as follows.

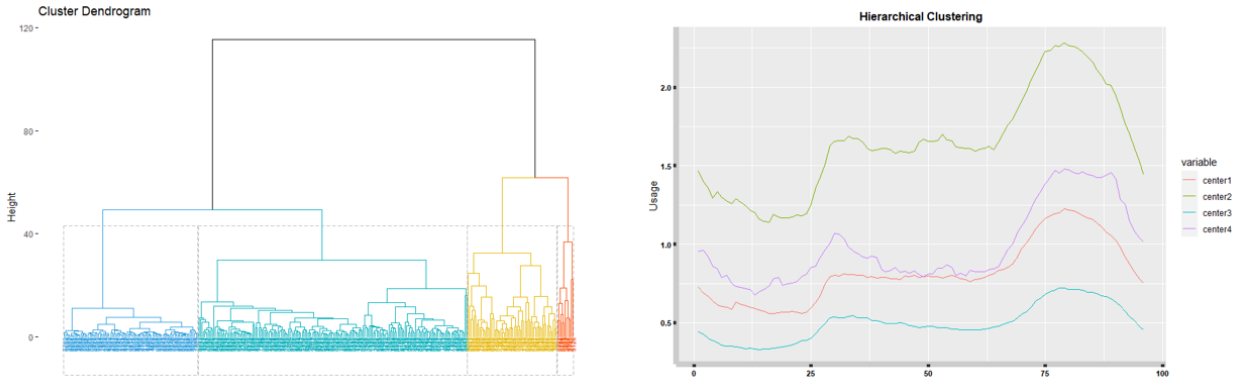



Figure 8: Results for Hierarchical Clustering

### Hierarchical K-means Clustering

Hierarchical K-means bargains the advantage of K-means algorithm in speed and hierarchical algorithm in precision. It is better used for the complex clustering cases with large numbers of data set and many dimensional attributes. The clustering result is as follows.

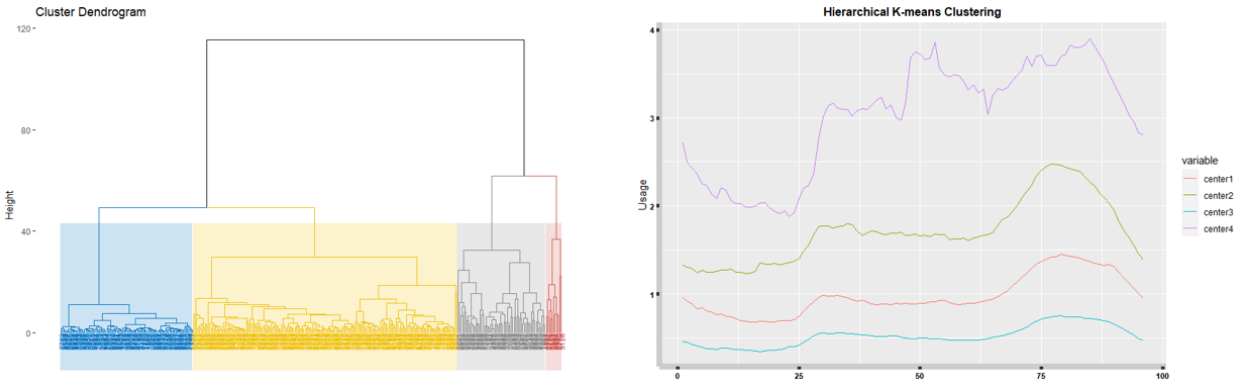



Figure 9: Results for Hierarchical K-means Clustering

### Model Based

Model-based clustering assumes that the data were generated by a model and tries to recover the original model from the data. The model that we recover from the data then defines clusters and an assignment of documents to clusters.

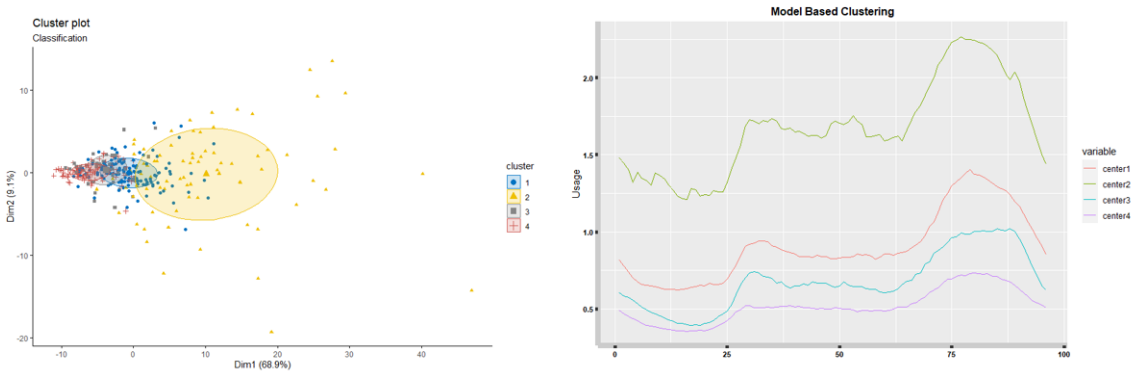



Figure 10: Results for Model Based Clustering

## 4.2 Analysis

### Performance Comparison

There are two internal evaluation methods, Average Silhouette Width and Dunn Index chosen to be the evaluation criterion for the clustering performance.

The average silhouette approach measures the quality of a clustering. That is, it determines how well each object lies within its cluster. A high average silhouette width indicates a good clustering. The average silhouette method computes the average silhouette of observations for different values of k.

The Dunn index (DI) is a metric for evaluating clustering algorithms.This is part of a group of validity indices including the Davies–Bouldin index or Silhouette index, in that it is an internal evaluation scheme, where the result is based on the clustered data itself.For a given assignment of clusters, a higher Dunn index indicates better clustering. The average silhouette and DI for each methods are listed in the following table.

| Criterion | Average Silhouette | Dunn Index |
|---|---|---|
| K-means | 0.3033794 | 0.03193218 |
| K-medoids | 0.2294664 | 0.02650151 |
| Soft K-means | 0.1276981 | 0.02525446 |
| Hierarchical Clustering | 0.2503529 | 0.03525853 |
| Hierarchical K-means Clustering | 0.3033794 | 0.03193218 |
| Model Based | -0.01193472 | 0.01394201 |

Table 1: Evaluation Criterion

According to the table, Hierarchical Clustering has the largest Dunn Index while K-means and Hierarchical K-means Clustering have the same highest average silhouette. Since for both evaluation criterion, the higher the index, the better the clustering, in addition to that the number of clusters is chosen according to the K-means, Hierarchical Clustering method is assumed to be the best method applied to pecan dataset.

### Usage Pattern Analysis

As shown in the Hierarchical Clustering result, It is shown that the four clusters have a similar electricity consumption pattern. From the beginning of the day, their consumption starts to descend and then increase around six in the morning. After reaching a consumption peak, it starts to fall with different level until five in the afternoon. Then the consumption starts to increase and reach its maximum of one day around eight in the evening. Then it keeps dropping until the day ends.

Hence it can be concluded that there are no extinct different consumption patters existing between different users other than consumption level. In other words, they have similar consumption trend instead of different usage patterns.

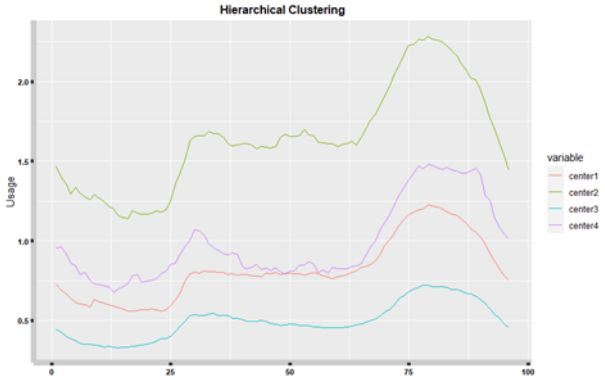



Figure 11: Hierarchical Clustering Centers

# 5 Weekdays and Weekends

As illustrated above, there are no extinct differences between the consumption patterns of different users. It is further explored that whether there exist different patterns among weekdays and weekends. In this case, the clustering was conducted within each day of the week. Set the number of clusters as 2, then the following results can be achieved.

According to the plots, the overall consumption trends are still the same, i.e. they all have two consumption peaks in the morning and evening respectively. However, an interesting fact show up that extinct different patterns exist between weekdays and weekends.

The first difference is about the appearance time of the first consumption peak. As marked with the red lines in the plots, the first peak appears around eight in the morning during weekdays (including Friday) while almost ten in the morning during weekends.

The second difference is about the performance after achieving the first consumption peak. During the weekdays (excluding Friday), there is a obvious drop of the electricity consumption after reaching the first peak at around eight in the morning. However, during the weekends (including Friday), the electricity consumption fluctuates around the peak value rather than decreasing to a relative low consumption level.

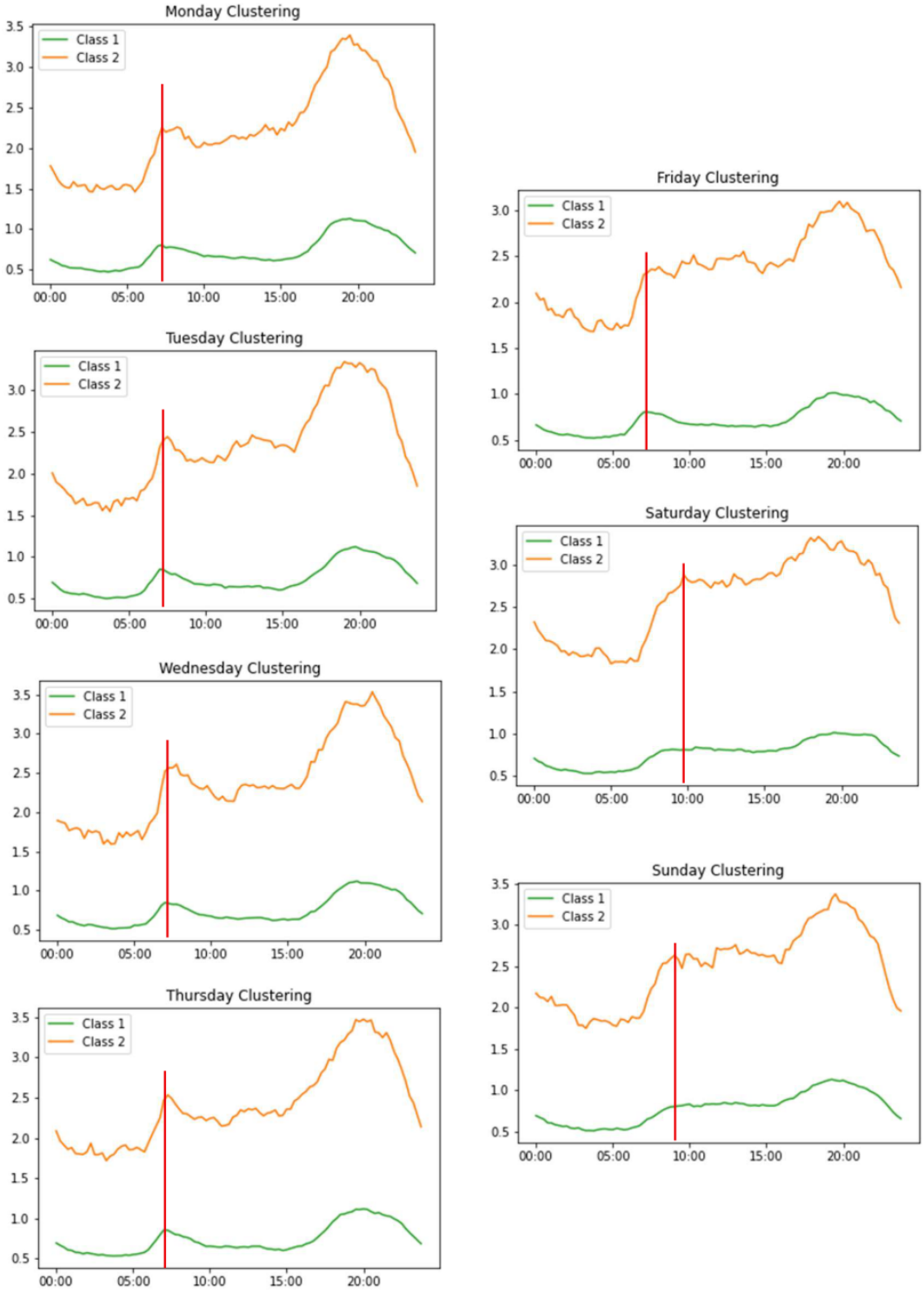



Figure 12: Clustering Among Weekdays and Weekends

On the other hand, electricity consumption during the daytime (around 10:00-15:00) is higher in Saturday and Sunday than those in weekdays, which is a main difference of patterns between weekdays and weekends. The assumption comes out that there may be more different consumption patterns during weekends. As a consequence, more patterns are explored for weekends.

To begin with, simply change the number of clustering to 3 and 4. The results are as following.

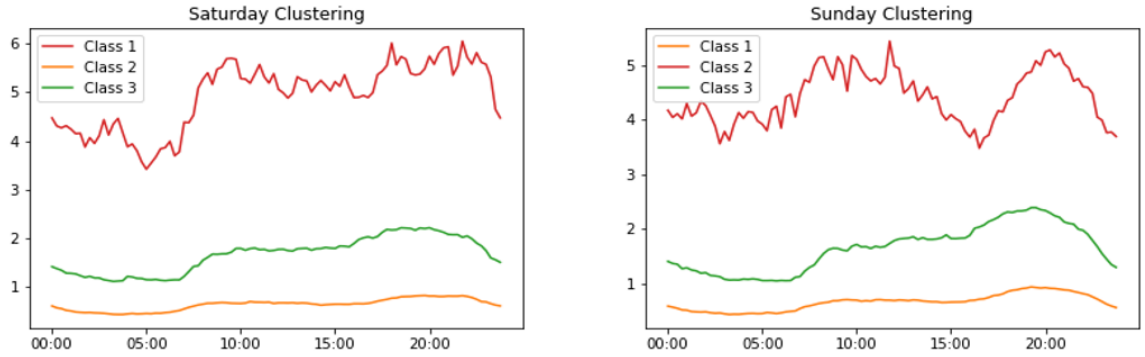



Figure 13: k=3

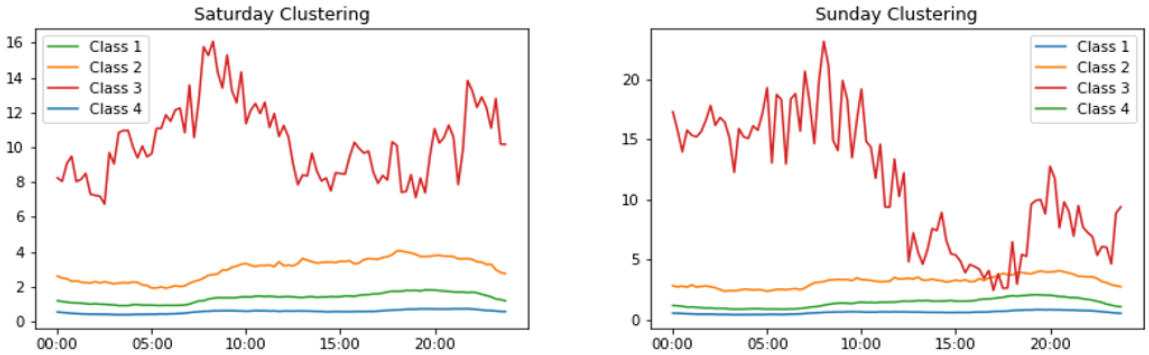



Figure 14: k=4

Some important rules can be detected from above figures:

1. When $k = 3$, the red consumption lines differ slightly from green and yellow ones, that is, the red consumption lines fluctuated drastically during the same time segment.

2. When $k = 4$, the red consumption lines differ much from green and yellow ones, especially during $15:00 - 20:00$. There is a abnormal decreasing trend for clusters in red during Sunday afternoon.

3. As the number of clusters increases, different consumption patterns began to appear.

From the figures above, it has been verified that various patterns are hidden behind the Saturday and Sunday, leading to a assumption that different patterns of electricity consumption will show up with respect to the increasing clustering number. However, it is time-consuming and hard to implement clustering algorithms with all possible number of clustering. Hence Shannon Entropy, a new concept reflecting randomness(or level of chaos, uncertainty) in patterns was introduced. It is a powerful tool for illustrating the problem.

# 6 Entropy

Entropy is a concept in information theory, which is also known as Shannon Entropy. In fact, entropy can be viewed as a measure of the uncertainty level of a random variable.

Recently, people have developed a formula to estimate the entropy of a sequence [16], which is shown as below.

$$S^{set} = (\frac{1}{n}\sum_{i}\Lambda_i)^{-1}\ln n$$

where $n$ is the length of the sequence, $\Lambda_i$ is the length of the shortest sub-string starting at position $i$ which does not appear previously from position 1 to $i-1$.

Through applying the sequence entropy formula to the electricity consumption data sequence, the value of entropy reflects the regularity of using pattern of a user. In consequence, the level of randomness in patterns during certain day (Saturday, Sunday $\cdots$) can be reflected by the distribution of all the entropy values computed from using data in that day.

Based on the meaning of the formula [17], it can be easily speculated that the small value of entropy represents a relatively regular using pattern of a user, and the large value of entropy represents a relatively chaotic using pattern of a user. Moreover, the day (Saturday, Sunday $\cdots$) contains more large entropy values tends to have a high level of randomness in using patterns among users, and the day contains more small entropy values tends to have a low level of randomness in using patterns among users.

The entropy values can be achieved via the following steps: for each day (Monday, $\cdots$ ,Sunday) and each user,

1) select all the using data in the day for the user

2) order the data by time, thus get a data sequence

3) apply the formula to compute the entropy value.

After these steps, the entropy value for each user on each day can be obtained. In this project, due to the limited computing resource, there are 185 users (346 in total) are selected to compute entropy, and each user has 7 entropy values (Monday, ... ,Sunday, respectively). Thus, there are 1295 entropy values which have already be computed. In order to view a rough distribution of entropy, a scatter plot is created as below.

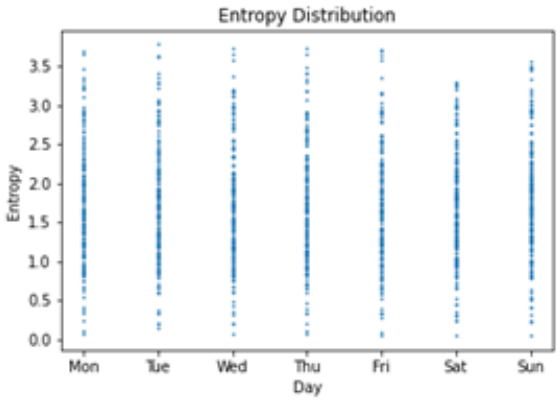



Figure 15: Entropy Distribution

The above graph is quite rough so that it can be found nothing special but a common discovery that the distributions in each day are similar. To get a better insight into this problem, the further work concentrates on the particular distribution within each day.

This part is mainly focused on Saturday and Sunday. First, cluster the using patterns among 185 users, label different patterns with different color. Then, plot the distribution

of entropy values from 185 users in certain day. Last, label the entropy value points with corresponding color represented in clustering result. The graph of results are as follows.

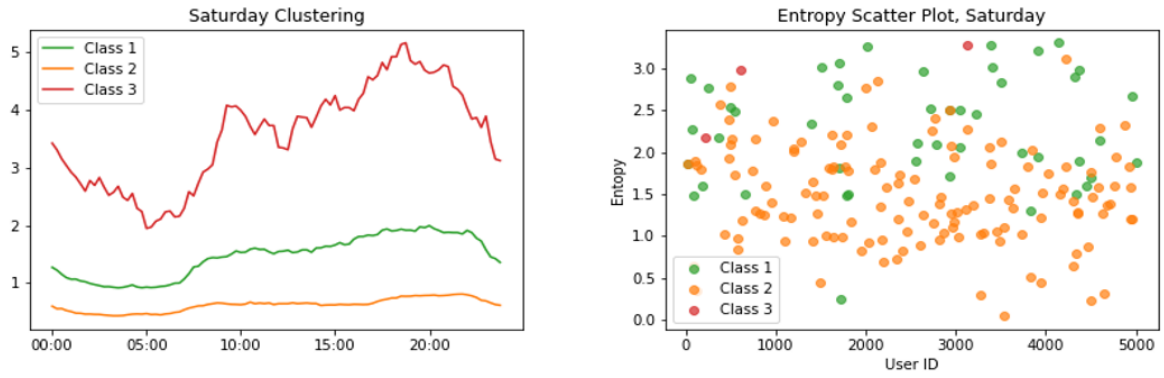



Figure 16: Saturday

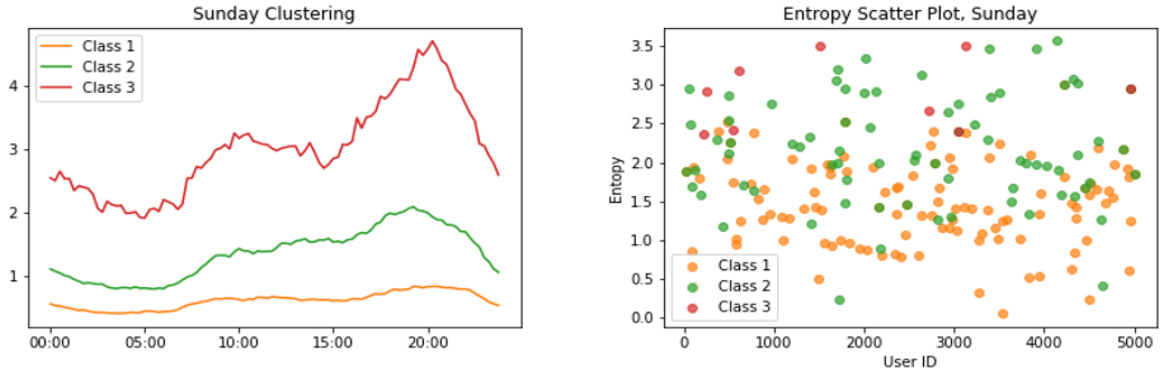



Figure 17: Sunday

Some important rules can be detected from above figures:

1. The entropy values in red color, which corresponds to a high consumption level, are distributed in a high position, revealing that users at a high level of electricity consumption use electricity in a more irregular way.

2. The entropy values in green color, which corresponds to a median consumption level, are distributed in a median position, revealing that users at a median level of electricity consumption use electricity in a common way.

3. The entropy values in red color, which corresponds to a high consumption level, are distributed in a high position, revealing that users at a high level of electricity consumption use electricity in a more irregular way.

Based on the above rules, it is possible to conclude that the irregular patterns of electricity consumption are responsible for the increasing of electricity consumption, on the contrary, the regular patterns of electricity consumption are the reasons for its decreasing. In terms of the electricity usage, suggestions can be brought up that following a regular electricity consumption pattern contributes to reduce the consumption.

This part mainly focuses on relating the entropy to the electricity consumption patterns in Saturday and Sunday. Future work can be expanded to more days in a week, for example, Monday, Tuesday, Wednesday, Thursday and Friday, which could bring up more insights about the regularity and randomness of electricity consumption patterns.

# 7 Conclusion

In this project, different clustering methods are applied to explore the electricity consumption patterns. According to Dunn Index, hierarchical clustering has the best performance on this specific dataset. Further, the similarities and differences of electricity patterns between weekdays and weekends are explored. It is found that a two main differences relating to the consumption peaks occur between weekdays and weekends. In addition, entropy was introduced to evaluate the randomness of the consumption patterns during weekends. It is discovered that following a regular electricity consumption pattern contributes to reduce the consumption.

# References


[1] L. Michelle, J. Rupamathi, and I. Marija, "Household energy prediction: Methods and applications for smarter grid design," IEEE, 2019.

[2] M. Krishnan, Y. M. Jung, and S. Yun, "Prediction of energy demand in smart grid using hybrid approach," IEEE, 2020.

[3] T. Sun, T. Zhang, Y. Teng, Z. Chen, and J. Fang, "Monthly electricity consumption forecasting method based on x12 and stl decomposition model in an integrated energy system," Hindawi, 2019.

[4] K. Muralitharan, R. Sakthivel, and R. Vishnuvarthan, "Monthly electricity consumption forecasting method based on x12 and stl decomposition model in an integrated energy system," Neurocomputing, 2018.

[5] M. Fayaz and D. Kim, "A prediction methodology of energy consumption based on deep extreme learning machine and comparative analysis in residential buildings," Electronics, 2018.

[6] J. D. RhodesaWesley, J. ColebCharles, R. UpshawcThomas, F. EdgarbdMichael, and E. Webbercd, "Clustering analysis of residential electricity demand profiles," Applied Energy, 2014.

[7] K. Karimi, "Clustering analysis of residential loads," 2016.

[8] Z. Zhang and T. Zimet, "K-means based clustering analysis of household energy consumption," 2019.

[9] Y. Huang, J. Zhan, L. W. Nana Wangand Chunjie Luo, and R. Ren, "Clustering residential electricity load curves via community detection in network," 2018.

[10] A. K. Zarabie, S. Lashkarbolooki, S. Das, K. Jhala, and A. Pahwa, "Load profile based electricity consumer clustering using affinity propagation," IEEE, 2019.

[11] M. Afzalan, F. Jazizadeh, and H. Eldardiry, "Two-stage building energy consumption clustering based on temporal and peak demand patterns," 2020.

[12] B. Dai, R. Wang, K. Zhu, J. Hao, and P. Wang, "A demand response scheme in smart grid with clustering of residential customers," IEEE, 2019.

[13] N. A.FundeMeera, M.Dhabu, A. Paramasivam, and P. S.Deshpande, "Motif-based association rule mining and clustering technique for determining energy usage patterns for smart meter data," vol. 46, Sustainable Cities and Society, 2019.

[14] "Pecan street." `https://www.pecanstreet.org/dataport/`. [Accessed: 2021-03-13].

[15] "Metadata view column descriptions." `https://www.pecanstreet.org/wp-content/uploads/2019/04/Dataport-metadata-dictionary.pdf`. [Accessed: 2021-03-13].

[16] I. Kontoyiannis, P. H. Algoet, Y. M. Suhov, and A. J. Wyner, "Nonparametric entropy estimation for stationary processes and random fields, with applications to english text," IEEE Transactions on Information Theory, 1998.

[17] C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," Science, 2010.