CptS 575 Data Science
Power Load Analysis Project Report
Brandon Bullard and Alexander Mantilla
12/11/21

Chelan PUD Substation Data Analysis Report

Table of Contents

0.0 Abstract

Electrical power is an incredibly important commodity in modern society. Most cultures are incredibly dependent on the availability of power and thus make power distribution an incredibly important field. In this study, we will investigate substation power data using a variety of data science, statistical, and mathematical techniques to try and find trends in power usage. With visualization techniques that include matrix profiling and exploratory data analysis, we succeed in gaining a stronger understanding of annual power usage and how different factors affect it. We ultimately find that stable power load levels do not necessarily equal stable power load level patterns.

1.0 Introduction

In the last 200 years, human civilization has begun to grow exponentially. One of the primary reasons for this proliferation of society has been the invention and use of electricity as a means of energy. Electrical power has allowed for the development of many technologies such as lighting, heating, and the internet. Widespread access to electricity has also allowed for the advancement of nearly all areas in science and technological development. Because electrical power is such an important aspect of human society, understanding the processes in which power is delivered from its sources to across the globe is crucial. The primary method for which electrical power is delivered to cities and towns is through electrical substations that store, convert, and distribute power to their surrounding area. These substations play a vital role in society because of their essentialness for power delivery in their surrounding area. With global increases in population, and the continued construction of new and more sophisticated buildings and towns, the demand for electrical power will only continue to increase. With this increase electrical substations will have to become better at delivering power. One solid way to improve substations is by optimizing their efficiency. When a substation is capable of storing and delivering the exact amount of power demanded it can be considered optimized. Optimizing the power output of a substation requires an analysis of the data that it collects. To perform an analysis on substation data, various techniques can be employed.

In this report, we will introduce a number of different analytical techniques that we use on substation power data obtained from Chelan county PUD in Washington State. We will look at the hourly power load data from ten substations and present different techniques that can be used to determine key insight into yearly power trends. By determining the rates of change of power load, and applying EDA and other statistical techniques such as matrix profiling to the data, we will not only find trends in past data, but also make numerical predictions for how power load will be demanded in the future.

2.0 Data

This section will explore the source, measurement, transformations, and distributions of the data for this analysis.

2.1 Description and Cleaning

The data in this paper comes from the Chelan County Public Utility District (CCPUD). Chelan county is located in central Washington, and maintains a population of about 80,000 residents (U.S. Census Bureau 2019). **Table 1** describes the variables provided by CCPUD. This data was comprised in two files, one containing all variables except for *total_district_load*, and the other file containing *total_district_load* with timestamps. The former file is hourly time series substation power and weather data, while the latter file simply aggregates the total power used by CCPUD for an hour. These two files are merged together on *timestamp_utcc* after data cleaning.

Table 1: Data Description

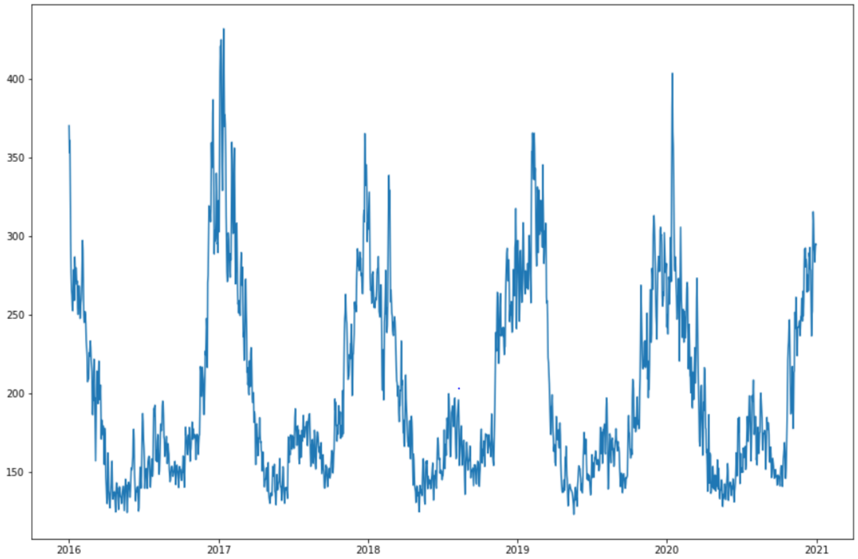
Variable	Description
timestamp_utc	ISO 8601 format
datetime_stamp	Pacific time format
sub_loc	10 Anonymized substation locations (a-j)
mw_hours	Energy used at that location during that hour
sub_category	Type of consumer (E.g. res = residential)
temp_loc	Temperature gauge location
temp_f	Temperature in Fahrenheit
total_district_load	Total energy use for Chelan county (mw hours)

After receiving data, it is important to understand if there are any errors and what the descriptive statistics are after cleaning has been conducted. The first step in the cleaning process was looking for null values. Thankfully, of the 438,481 hourly observations only 18,000 had missing values for *mw_hours* and *temp_f*. Since these missing values can bias the estimates for *total_district_load*, it was our decision to remove the missing values rather than impute them with the means during other years. Outliers were the next focus, and there were a handful on both the high and low extremes. There were 6 observations where there were unreasonably low *temp_f* values, and approximately 100 observations that had *mw_hours* in the thousands. These outliers were removed from the dataset. Values for *mw_hours* below .1 were also excluded, as these were determined to be data errors as well. The data spans from January 1st 2016 to the beginning of 2021. For a more consistent look year over year, the partial data for 2021 is not included in the analysis. With the data finally in a clean and usable form, exploratory data analysis is the logical next step.

2.2 Exploratory Data Analysis

Before choosing analytical methods, it is important to understand the nature of the data. In order to understand how energy use varies over time, **Figure 1** illustrates how *total_district_load* varies over time.

Figure 1: Power Usage Over Time



It is clear that there are seasonality trends in the data; winter looks like it has the greatest energy demand while warmer months are consistently lower. Summer months like June and July can be perceived as the smaller peaks between the large winter demands. Given the spatial data provided by CCPUD, a closer look at local trends can also inform analysis efforts.

Figure 2: Total Power Use and Temperature

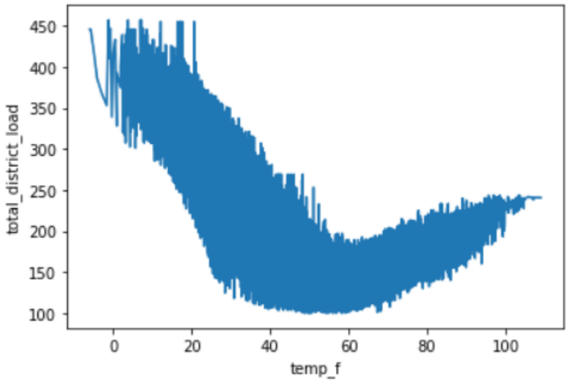


Figure 3: Total Power Use by Year

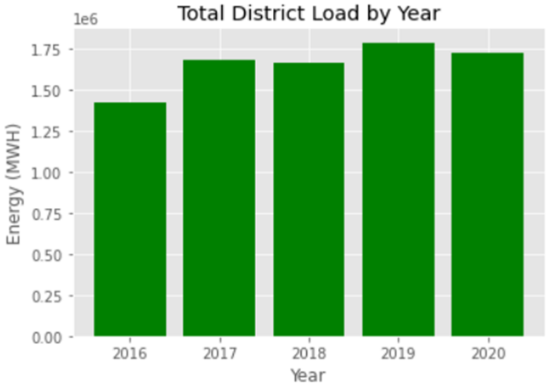


Figure 2 demonstrates how total power use changes with respect to temperature. From this illustration it is clear that temperature and power use do not have a directly linear relationship. It appears that lower temperatures place a much larger demand on power, middling temperatures place the least demand on power, and higher temperatures demand slightly more. In order to control this relationship, *temp_f* could be transformed to capture degrees above and below low power “comfortable” temperatures. Dummy variables for seasonality might also capture this effect, but additional dummy variables could be created at temperature boundaries. Figure x captures how overall power use has changed through each year. While power use has increased from 2016 to 2020, this trend does not consistently hold year over year. In fact, after total power use increases one year it slightly decreases the next, only to surpass the previous high point. Overall, the trend is positive which is important to be mindful of as time series data is already susceptible to collinearity.

3.0 Analysis Methods & Models

3.1 Total Load and its First and Second Order Time Derivatives

One of our primary methods of analysis will be taking a visual inspection of the hourly total district load, its rate of change, and its rate of rate of change from 2016 to 2021. We are specifically interested in determining these orders of rate of change because we believe they can provide information as to what times of the year tend to have the most stable power load, and what times of the year have the most power load changes. To determine the rate of change or 1st order derivative of the total district load, we calculated the slope between each hourly data point using the formula:

$$\frac{dP}{dt}\left(t+1\right) = \frac{P(t+1) - P(t)}{\Delta t}$$

where P is the total district power load at a given time t. Using this relationship we were able to generate a matrix of rate of change observations for the total district load. We used the same process on the rate of change data to find the rate of the rate of change data or 2nd order derivative:

$$\frac{d^2P}{dt^2}\left(t+1\right) = \frac{\frac{dP}{dt}\left(t+1\right) - \frac{dP}{dt}\left(t\right)}{\Delta t}$$

.

These equations were implemented into code in R by specifying t+1 to be the next interval time step collected (which was equal to 1 hour), and by specifying Δt to be the difference between t+1 and t. In our case, Δt was always equal to 1 hour or 1 in our code.

3.2 Developing a Matrix Profile

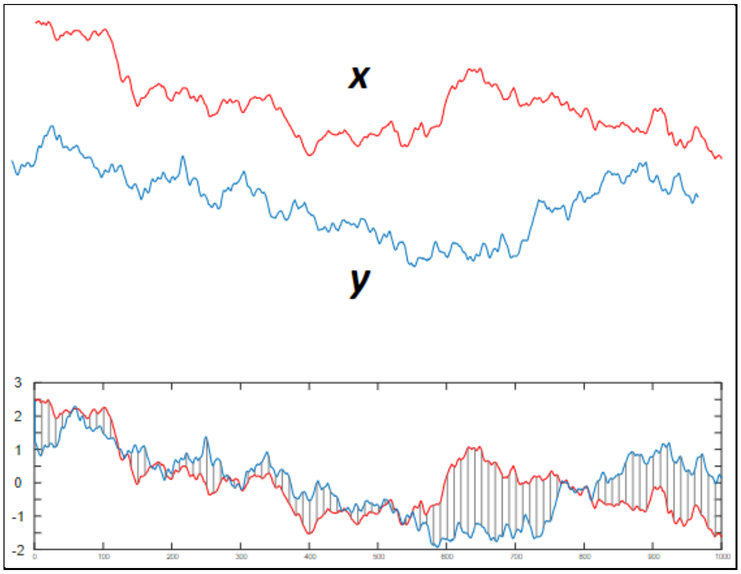
Another primary analysis method that we employed was the use of a matrix profile for the total load as well as its 1st and 2nd order derivatives. Because of the way power grids distribute electricity, it must be generated at the time energy is demanded (Hitchen 2017). Therefore, accurate anomaly detection methods are important for anticipating demand so energy can be generated. The matrix profile analysis method was once again a visual heavy method, and was used to determine how often total power load patterns repeated themselves between 2016 and 2021. The idea behind the matrix profile was to cut the total hourly data into small pieces and compare each piece or sequence to the others. By comparing each sequence, we could develop an idea for whether a sequence appeared to repeat itself if it had similar sequences. Before we compared our sequences of the data to each other, we normalized all sequences to prevent any sort of drift between sequences. Normalization was done by obeying the equation:

$$\hat{x}_i = \frac{x_i - \mu_x}{\sigma_x}$$ [2]

where each normalized sequence $\hat{x}_i$ is determined by subtracting each sequence x_i by the average value μ_x of its included data points and dividing it all by its standard deviation σ_x. With normalized sequences, we then make a comparison between every sequence with every other sequence by determining the Euclidean distance. The Euclidean distance represents the total distance between each data point of two sequences, and is defined as follows:

$$d(x,y) = \sqrt{\sum_{i=1}^{n} (\hat{x}_i - \hat{y}_i)^2}$$ [2].

where distance d is determined by looking at the total magnitude difference between each consecutive data point i between two sequences $\hat{x}_i$ and $\hat{y}_i$ This distance gives us a direct quantification as to how similar every sequence is to each other. A visual representation of the Euclidean distance is shown below.



source: https://www.cs.ucr.edu/~eamonn/Matrix_Profile_Tutorial_Part1.pdf

To create a matrix profile with Euclidean distances, we saved only the minimum distance value from every sequence and recorded that value at the given sequence's time in the total data. This gave us a quantification of how repeating different points in time were between 2016 and 2021.

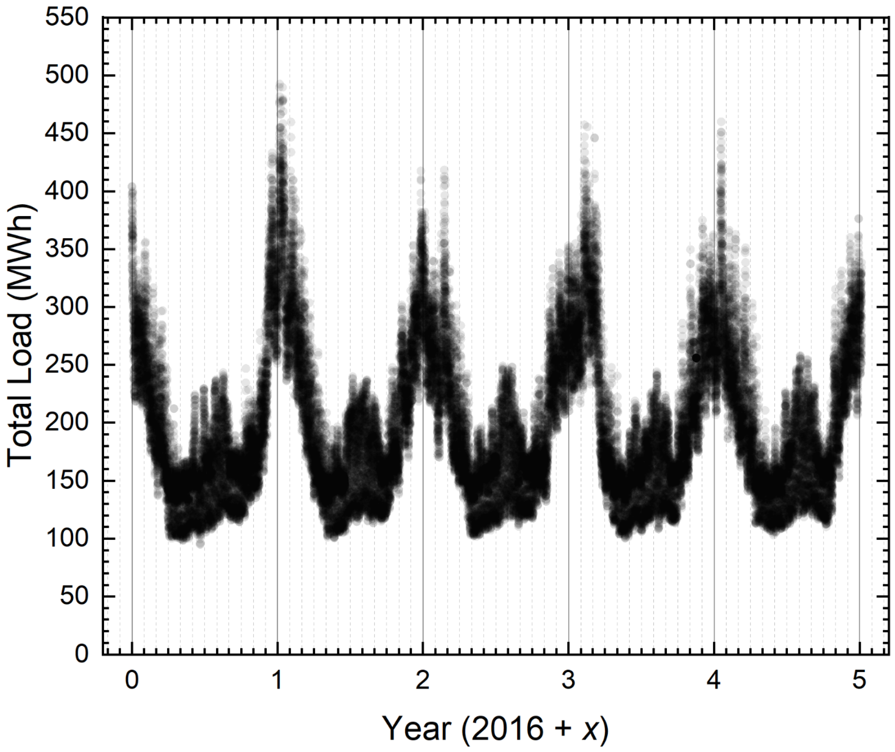
The actual processing time for a matrix profile can be extremely long. To avoid this, we made a few key decisions with how we recorded our matrix profile. The first decision was that we would use 6-hr sequences. By having 4 sequences per day, we would be able to extract accurate information about how repeated power sequences were, without having incredibly long processing time. Our second decision was to make comparisons between uncorrelated sequences only. To ensure sequences were uncorrelated, we ensured that data points in each sequence were only used once. Each sequence contained its own set of data points and no two sequences shared data points.

4.0 Results & Discussion

4.1 Rate of Change Analysis Results

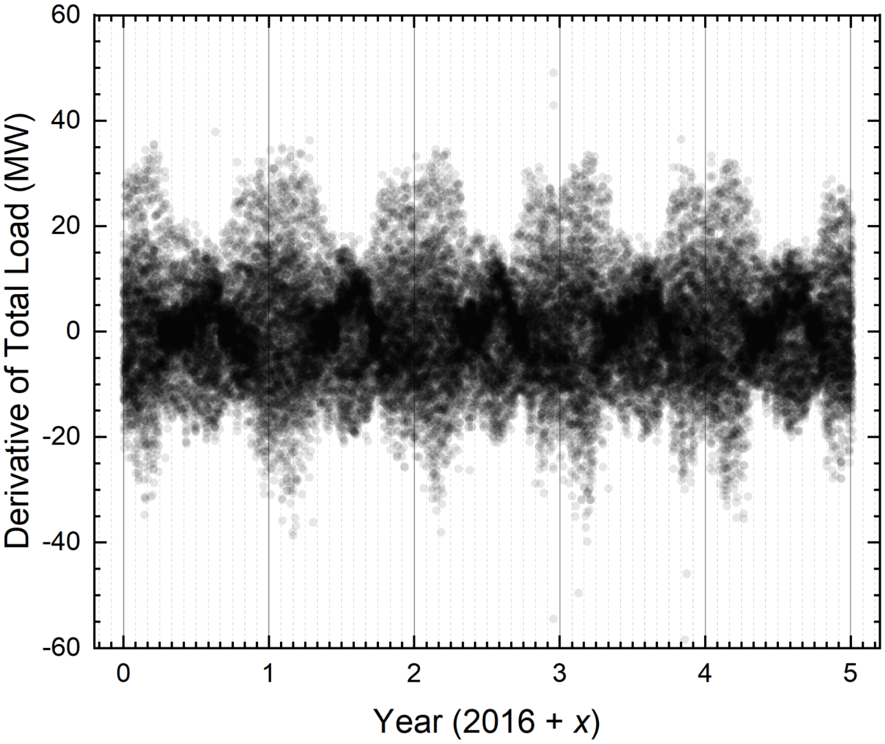
One of the primary studies we wanted to conduct was determining both expected and unexpected phenomena in the total power load. This goal was in the interest of better predicting patterns and thus optimizing the substation. The 1st derivative of total load was determined by finding the slope between each point in the total load and plotting all discovered slopes. This collected data frame represented the rate of change of the total power used per hour. A negative value represents the power load decreasing, and a positive value represents the power load increasing. Next, the 2nd derivative of total power load was determined by finding the slope between each data point of the first derivative. This collected data frame represents the rate of change in which the rate of change of power use per hour changes. In this case, a negative value in the second derivative represents a decrease in power load being decreased (power load decreases slower), and a positive value represents an increase in power load being increased (power load increases faster). We decided to look at the change between each hour to get the highest resolution of data possible for the 1st and 2nd derivative. **Figure 4** shows the calculated total load from 2016 to 2021, as well as its rate of change for every hour:

Figure 4: Total Load vs. Time



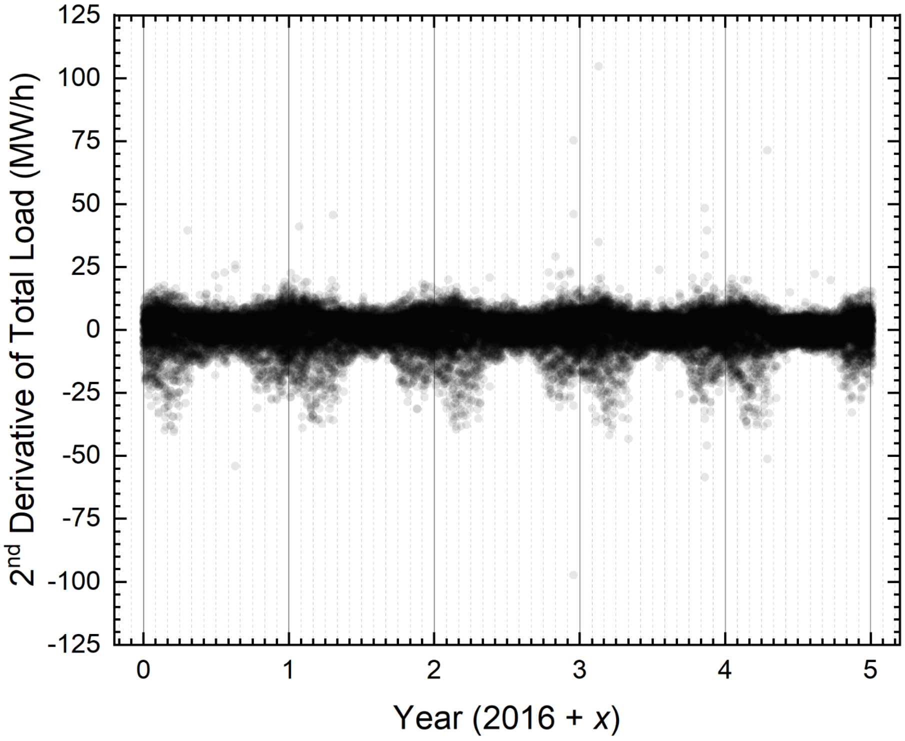
Total power load primarily stays between 100 and 500 MegaWatt hours. At the beginning of each year, total power load reaches its peak value between 400 and 500 MWh. There are a few reasons why the power peaks around this time of year. The temperature drop from winter is one of the larger factors. When the temperature drops in the winter the use of electric heaters increases. This increase causes the total power load to increase. Alongside temperature, the winter has shorter days and longer nights. These longer nights cause electrical lighting to be a larger factor in power drain. Also, major holidays such as Christmas and New Years play large factors into power drain, causing the total drain to be much larger in the winter. Following the start of the new year, the total load once again begins to drop. This drop likely correlates well with temperature and reaches a yearly minimum between March and May. After May and into the summer, the power load begins to slightly increase again between mid June and early September. This slight increase is likely attributed to temperatures being high enough to warrant the widespread use of air conditioning. After September, the power load once again increases into the new year and the process repeats itself. Since power usage is higher in the winter months, power providers should expect to have highest demand for power between December and February of every year. Power providers should also expect a slight increase in power demand in the late summer.

Figure 5: 1st Derivative of Total Load vs. Time



Our rate of change of total district power load tells a more complicated story. As seen in **Figure 5**, the total rate of change of power load always stays between -60 and 60 MW. For the purpose of this study, we will refer to any stretches in time with changes to power load above ±20 MW as *power fluctuation periods*, and any stretches in time with changes in power load below ±10 MW as *power equilibrium periods*. Between early 2016 and early 2021 we see 6 power fluctuation periods and 5 power equilibrium periods. All fluctuation periods appear to reside between the end of September and the start of May of each year, and all equilibrium periods reside between the beginning of May and the end of September. The fluctuation periods correlate with the period of the year where temperatures are lower due to winter, and the equilibrium periods correlate to summertime. Another finding that is a bit harder to see is a trend for the consistent increase in power demand every June to September. Using partially transparent data points, a consistent increase in power load can be seen in **Figure 5** starting at around 0 MW in June and increasing to 10 MW in August, then back down to 0 MW in September. This phenomenon appears to be the most common feature every summer as it is darkest, and suggests that a consistent demand of 10 MW is found every August. Both these findings suggest two primary features for power providers to look for. Since it appears that large (above ±20 MW) increases or decreases in power use only happen during the winter, it can be fairly confidently stated that power usage in the winter is both higher than the summer and more dynamic. Also, power demand increases at about 10 MW every August, which power providers should also monitor and prepare for on a yearly basis.

Figure 6: 2nd Derivative of Total Load vs. Time

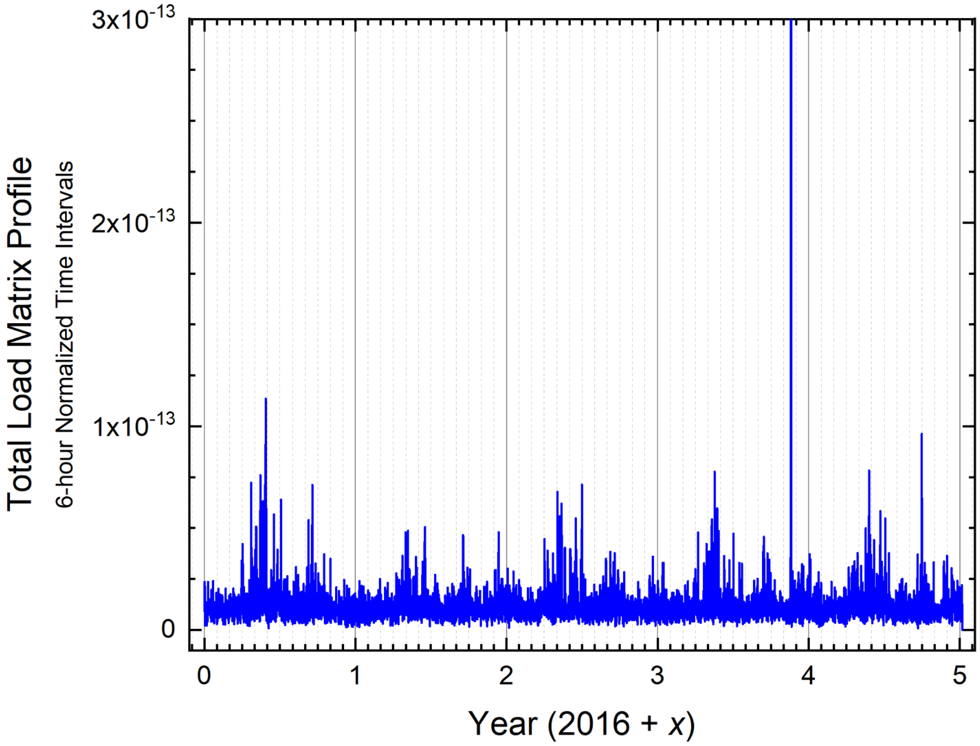


Our rate of change of the rate of change for total district power load tells a similar story to the 1st derivative shown in **Figure 6**. The figure shows that power demand increases and decreases tend to never increase or decrease more than ±50 MW/h. This means power demand does not increase or decrease by more than 50 MW every hour with a few exceptions. The nature of these exceptions is unknown, with no more than a few occurring in the past 5 years. One possibility is that these extremely high changes in power load demand changes are attributed to outages or substations quickly turning on. Aside from the outliers in the 2nd derivative data, There appear to be 6 power fluctuation periods and 5 power equilibrium periods. Similar to the 1st derivative data, the 2nd derivative shows consistent changes to power load change between September and May of every year. However unlike the rate of change of power load data, the rate of change of power load rate of change decreases at much higher rates than it increases. In the winter, the total power load change appears to only increase by 15 MW/h but at times decrease by upwards of -40 MW/h. This may show a systematic limitation of the substations that regulate this power, in that substations can cut off power much faster than they can provide power. Aside from the fluctuation regions, the summer months of May to September show a power equilibrium. In these months, the rate of total power load change almost never reaches above ±10 MW/h, suggesting that power providers can expect a steady demand power with potential demand increases that never increase faster than 10 MW/h.

4.2 Matrix Profile Analysis

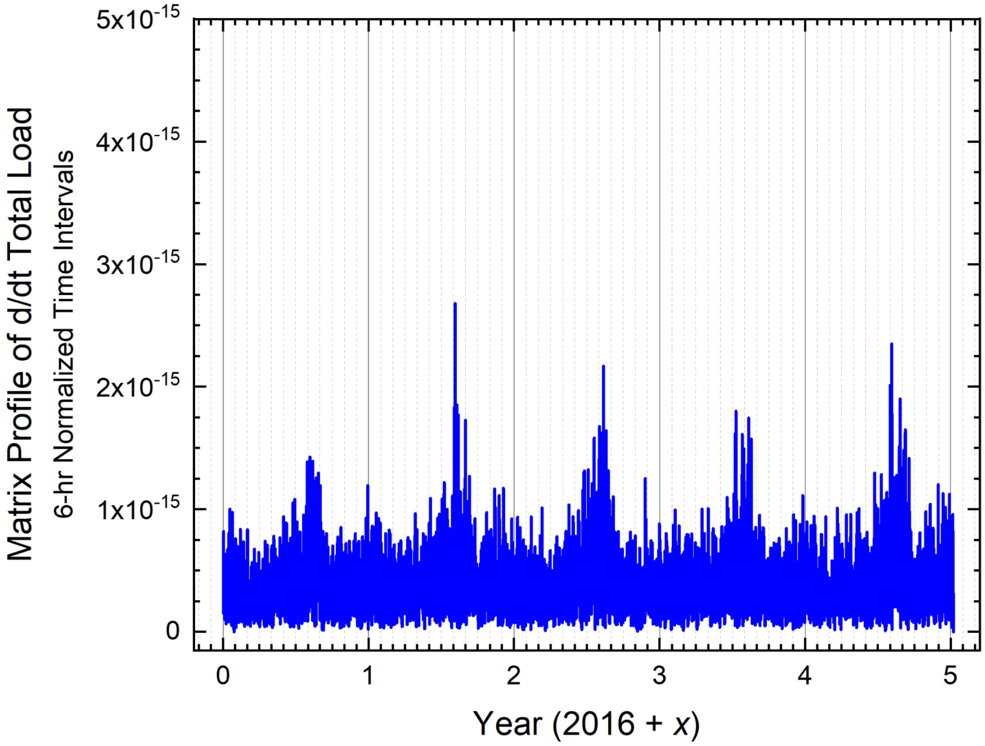
Beyond looking purely at visual patterns in the total load and its 1st and 2nd order derivative, we also wanted to be able to visualize both anomalies and motifs in the data from 2016 to 2021. We wanted to be able to show when and where power features were repeated as well as when they were unique. To accomplish this analysis, we separated all of the hourly data into ~7000 different 6 hour sequences. We used the Euclidean distance formula to look at the distance between each given sequence and all other sequences in the data. Before the sequences were compared, they were normalized to prevent skewed data values from data drift. The minimum detected distance was stored and placed at the time location of the sequence to specifically show how close the closest sequence was to the given sequence. Lower values on the profile represented sequences that had very similar sequences at some other point in the data. Higher values on the profile represented sequences that were less common, and likely affected by anomalies rather than patterns in the power data. This technique was not only done on the total load, but also its 1st and 2nd order derivatives. The matrix profile for the total power load is shown on **Figure 7**.

Figure 7: Matrix Profile of Total Load vs. Time



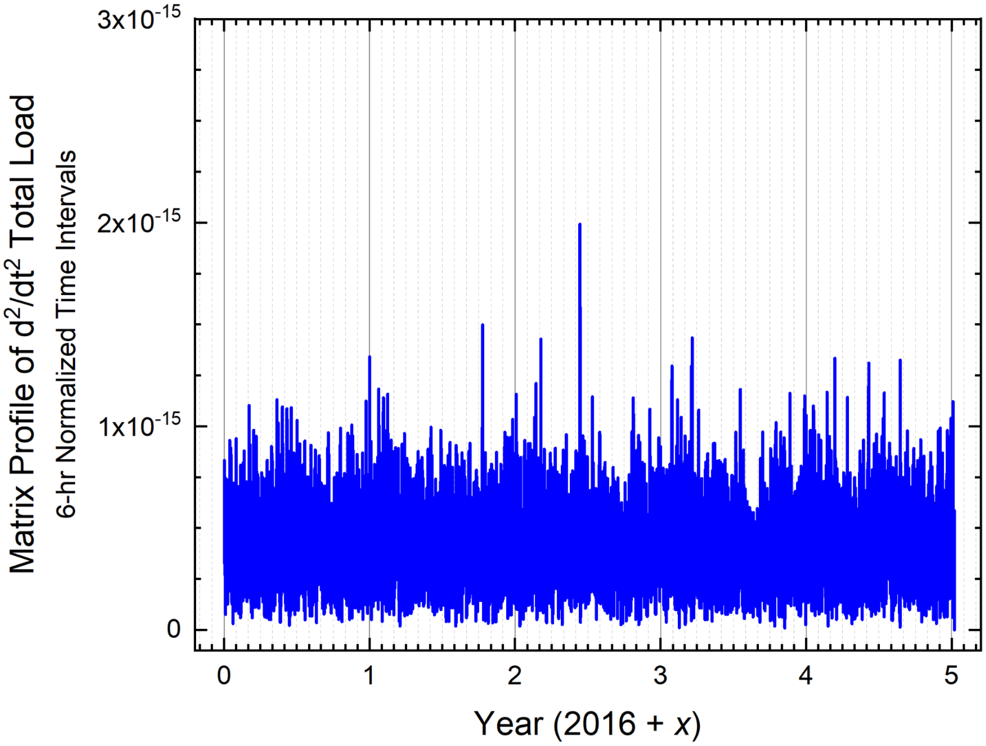
The matrix profile of total district load shows very few anomalies in the total load. For all 7000 of the 6-hr sequences the minimum distance appears to be on the order of 10^{-13}. This suggests that for every 6-hour period between 2016 and 2021, the total power load has at least one match that is extremely close to it. However, while pretty much all sequences appear to have similar matches, some have stronger matches than others. The majority of sequences appear to have closest matches between 0 and $2x10^{-14}$. These sequences most consistently appear in the winters of each year. In the early summers of each year, especially around the beginning of May, the matches for many sequences rise to as high as $1x10^{-13}$. This consistent rise suggests that power drains are consistently more volatile during this time of year, while the winter appears to be more patterned and predictable. There is also a significant peak that arises at around November 18th, 2019. The reason for this feature is unknown, but for whatever reason, the power features in this sequence rise to around $2x10^{-12}$, significantly higher than the rest of the data and suggesting that a power anomaly occurs on that day.

Figure 8: Matrix Profile of 1st Derivative Total Load vs. Time



The matrix profile for the rate of change of total power load shows interesting periodic phenomena. As seen in **Figure 8**, the sequences made in the rate of change matrix profile have much closer matching sequences. The Euclidean distance of most normalized sequences resides between 0 and $1x10^{-15}$, which is a factor of 100 smaller than the distances of the total load sequences. This suggests that the rate of change of power load on a 6-hr basis is a bit more predictable. However, there are 5 notable spikes in Euclidean distance in the figure that all happen to occur between August and September of each year. In these spikes, the distance rises to between 1.5x10^{-15} and 3x10^{-15}. This finding suggests that power load rates remain fairly steady throughout the year, but can be found to be a bit more sporadic near the end of the summer.

Figure 9: Matrix Profile of 2nd Derivative Total Load vs. Time



The matrix profile for the 2nd order derivative of total power is much less interesting. When looking at **Figure 9**, almost all sequences remain below 1x10-^{15} with only a few sequences above. For the most part, none of the larger distance sequences seem to follow any sort of pattern. The largest distance sequence is around June 9th, 2018. Around this day, the change in the rate of change of power load was unusually different than any other day in the last 5 years. The reason for this difference is unknown, but interesting regardless. The results of this matrix profile suggest that the 2nd order derivative of the total district rate of change does not have many anomalous features, and can be predicted very accurately from past data.

5.0 Conclusions

From our analysis of power load data in Chelan county from 2016 to 2021 we were successful in developing a deeper and more intimate understanding of how power load varies with different governing factors. Using a variety of different data science and mathematical techniques, we were able to draw interesting conclusions about how temperature, time, and place can all affect power load.

EDA proved useful in developing intuition about the data and how energy use varies. This was a key step before modeling as many statistical methods work best when they are chosen with respect to data (E.g. gathering count data and using Poission regression). Foundational questions were also answered by simply plotting and observing relationships. For instance, energy use does not follow a simple linear relationship with respect to temperature, and the growth of total energy use varies year over year.

From our study of the 1st and 2nd order derivatives of total district load, we found that power load fluctuates throughout the winter months of the year and equilibrates in the summer. We have predicted that temperature and season is a large factor in whether or not power load is stable. Through the 2nd derivative we have also found that substations are able to decrease their power load much quicker than they can increase. Matrix profile analysis revealed a different story. Profiles of total load and its rate of change revealed that the end of the summer has a higher rate of power load anomalies each year, while the winter appears to have more consistent power loads on a yearly basis. Profiles from all three datasets even revealed a few large anomalies that we have yet to identify. The combination of both pieces of data suggested an interesting conclusion: Power load is more volatile and changes at higher magnitudes in the winter, yet these changes are more predictable on a yearly basis than changes found in the summer. While this conclusion was an unexpected result, we justify its validity by considering that we found the conclusion to be consistent at all levels of the analysis. From our findings, power load changes at higher and faster rates in the winter, yet these changes are more predictable than the lower and slower changes in the summer.

6.0 References

[1] Hitchen, Penny. “Energy Demand Forecasting in a Rapidly Changing Landscape.” GE, Dec. 2017,https://www.ge.com/power/transform/article.transform.articles.2017.dec.energy-demand-forecasting-in-a

[2] Mueen, Abdullah. “Matrix Profile Tutorial”, www.cs.ucr.edu/~eamonn/MatrixProfile.html

7.0 Appendix

Corresponding code for this project can be found here:
https://github.com/branbull/CptS-575-Proj