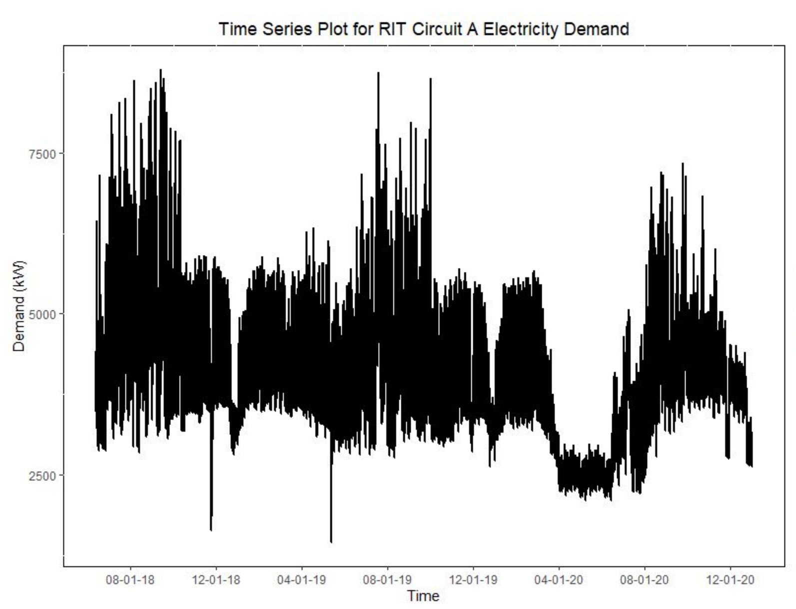
RIT Energy Consumption Data Cleansing

Trevor Martin
02/25/2021

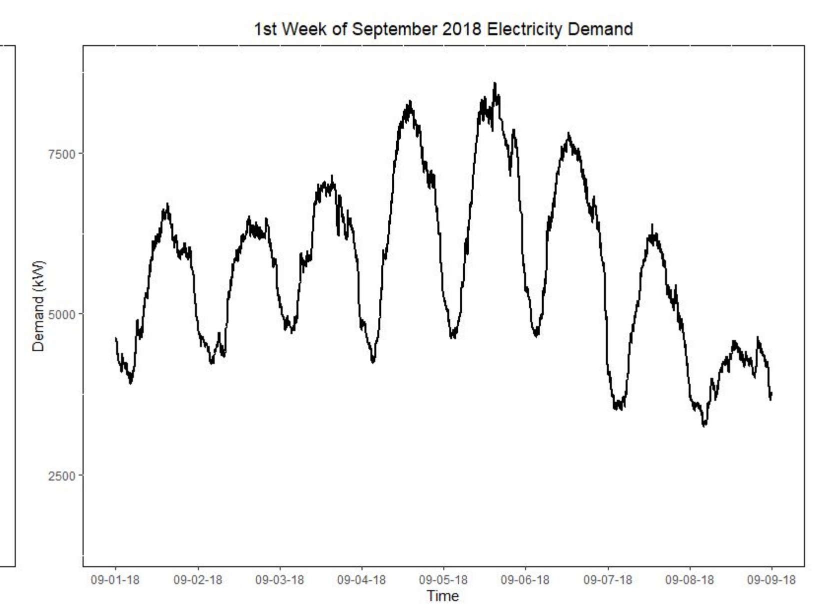
Introduction

RIT consumes a large amount of energy every day. Predicting this usage can be useful to mitigate risk of over consumption, which can lead to large electricity bills with demand charges accounting for up to 70% of any given bill. Demand response activities can be completed if a peak day is predicted to lower the demand charges RIT faces.

RIT Electricity Demand

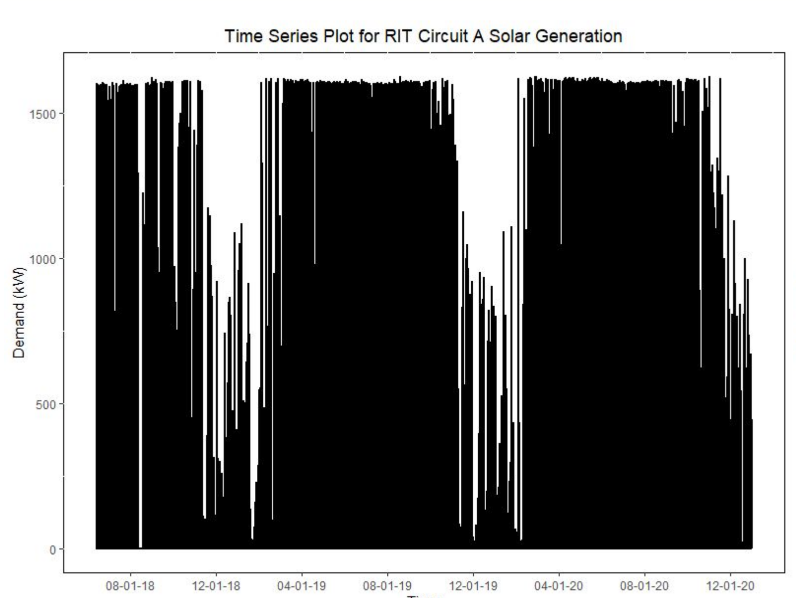


Demand between 06-13-2018 & 12-31-2020 in 5 minute intervals. All subsequent plots use 5 minute interval data.

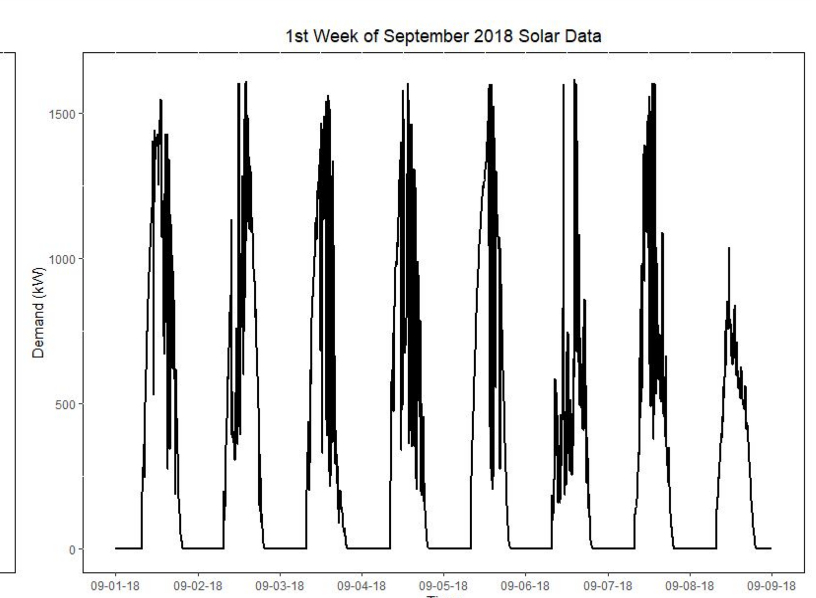


Shows the first week of 09/2018 and the seasonality that is present each day for electricity demand. Demand is lower at night, and higher during the day.

RIT Solar Generation

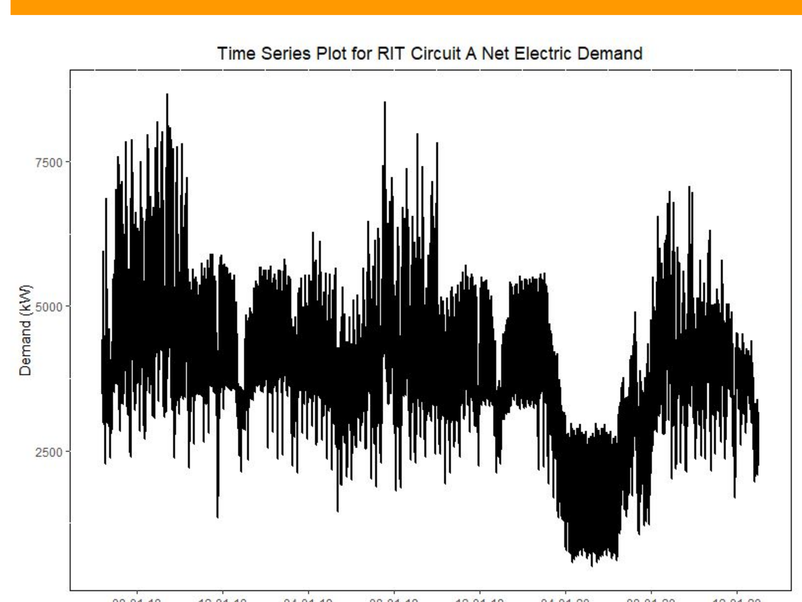


Solar generation between 06-13-2018 & 12-31-2020. It can be seen that winter has lower solar generation than the rest of the year.

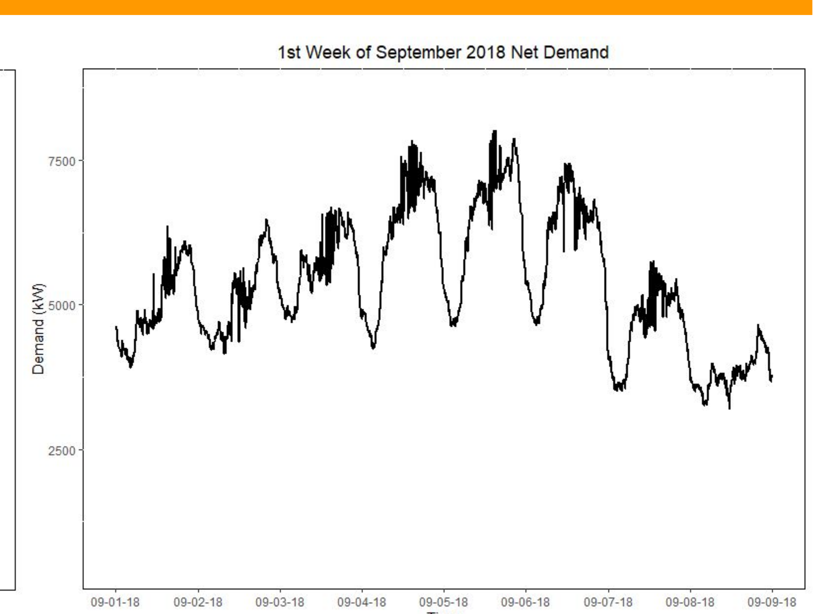


Shows the first week of 09/2018 and the seasonality that is present each day for solar generation. Solar generation is high during the day and zero at night.

RIT Net Demand



Net demand for electricity by RIT. This is calculated by subtracting the solar generation from the electricity demand. This data is used for the majority of the analysis going forward.



This plot shows the seasonality present each day with net demand. The peaks of this are lower than demand, due to the peaks of demand being aligned with the peaks of solar generation.

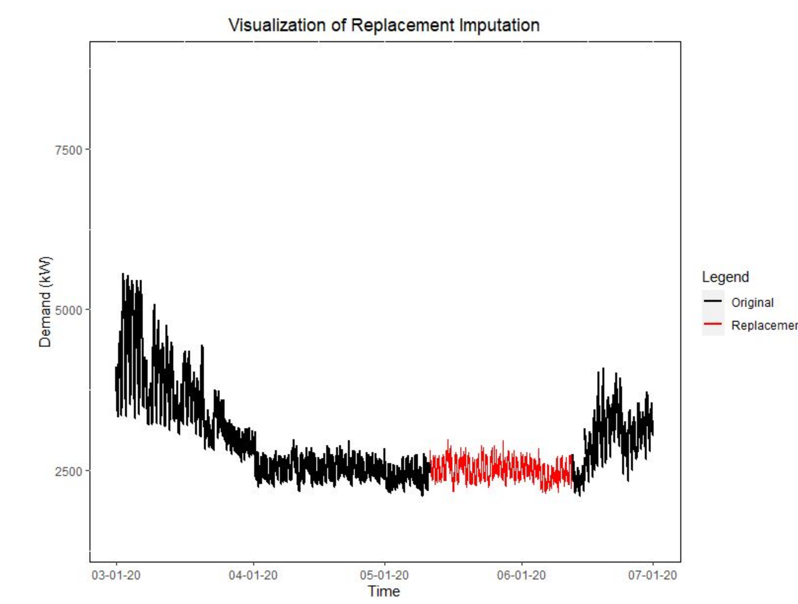
Data Cleansing

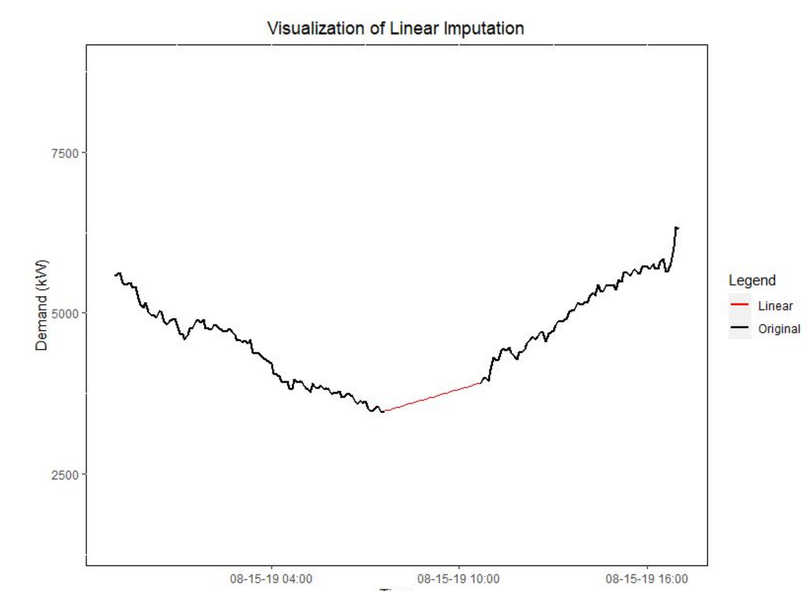
This table shows the number of missing or anomalous values present in the raw data.

Dataset	Count of missing or anomalous values present	Percent of all values
Demand	9849	3.66%
Solar	9681	3.60%

Imputation

The first imputation method used was utilizing historic data to replace missing values, which is highlighted in this figure. The missing data chunk started on a Monday, so a similar sized data chunk was taken with a Monday start to fill in this missing data. This method was used for the highlighted chunk and one other large chunk of missing data.





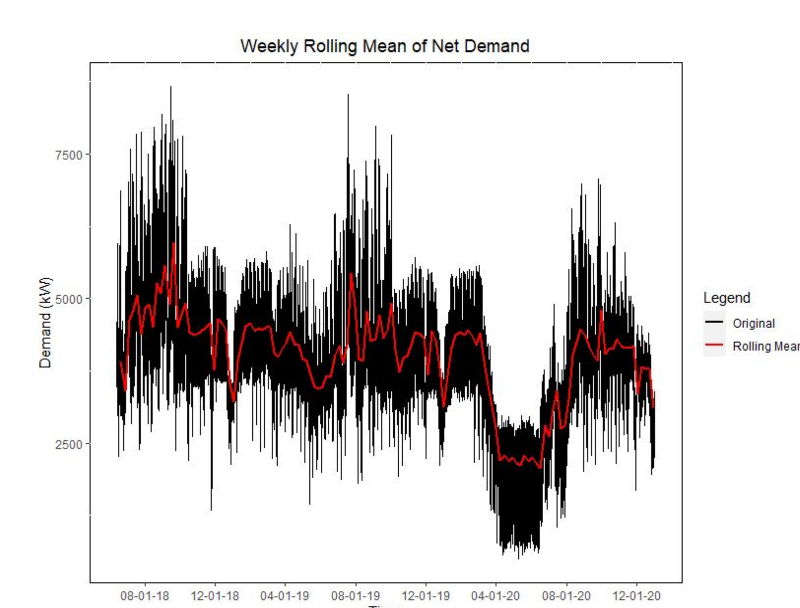
The other imputation method used was linear interpolation. This was used on many small gaps to fill in missing data points. An example of this is highlighted in this figure. This linear method was also used for all of the solar data except for one large gap where historic replacement was used.

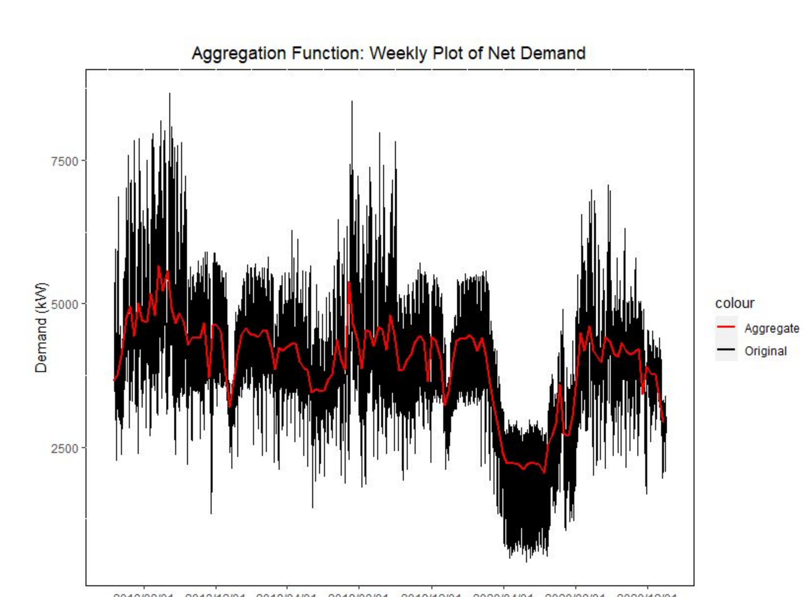
Statistics & Smoothing

Dataset	Mean	Variance
Demand	4220.003	1135917
Solar	234.590	180073
Net Demand	3985.412	1093263

After cleaning the data, these summary statistics were calculated for each of the 3 data sets. Mean is in kW and variance is in kW^2.

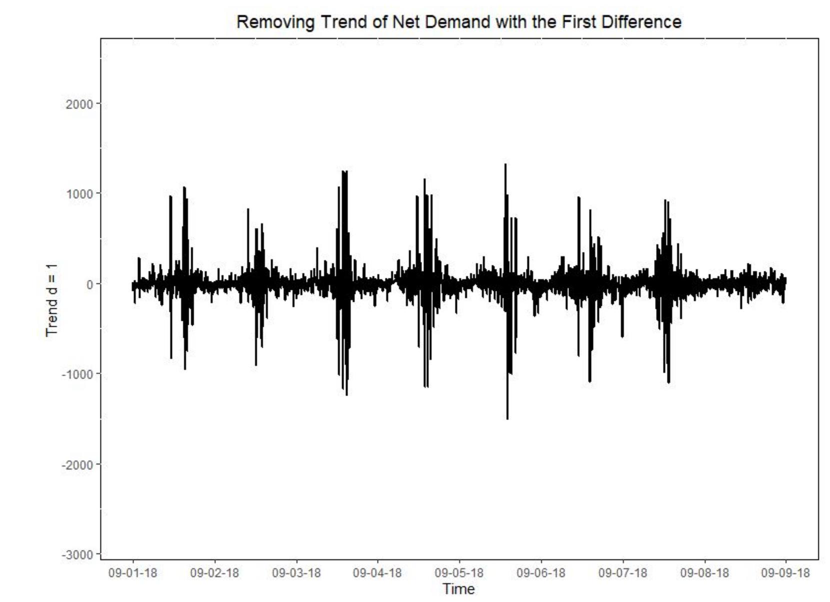
The weekly rolling mean is used to smooth the data. This plots the unweighted average of the last week worth of data. This can be useful to get an average interpretation of the data rather than looking at all of the raw data.



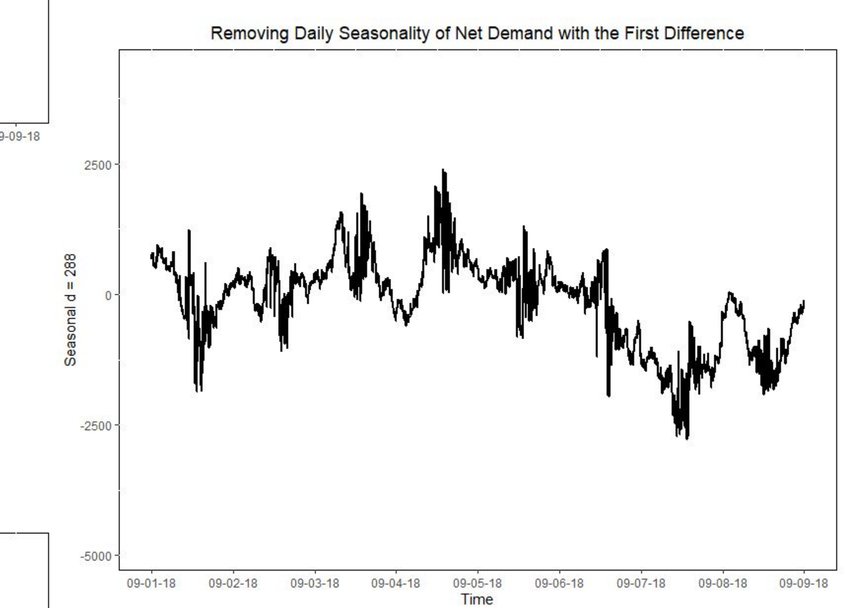


The weekly aggregate function acts similarly to the rolling mean, except it calculates the mean of the week ahead of the current point. As a result of this, it plots a week behind the rolling mean.

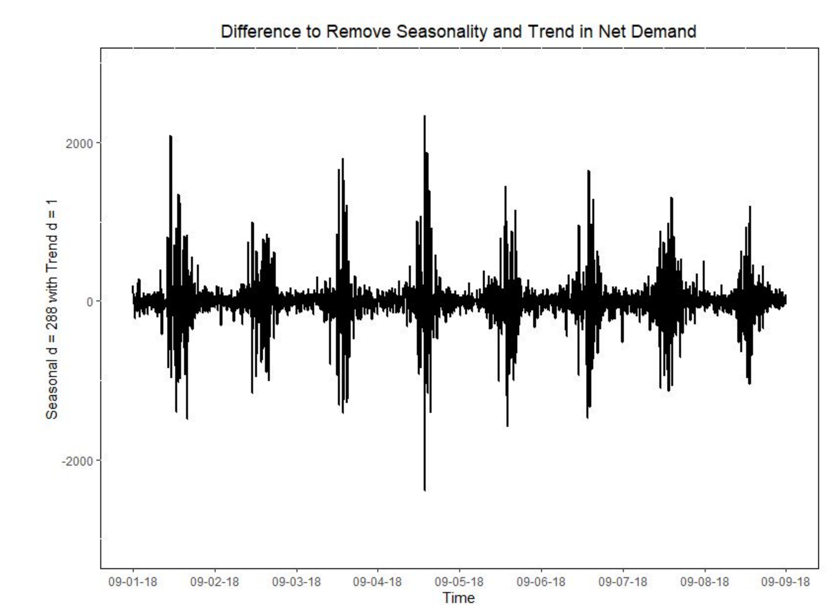
Differencing to Remove Trend and/or Seasonality



Detrending the data gives a mean of zero, but a non constant variance.



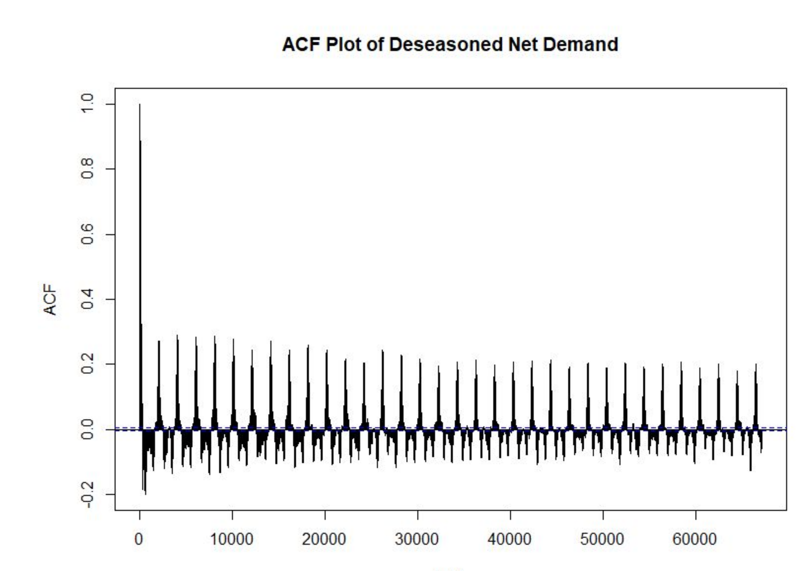
Deseasoning the data gives a mean of zero, but a non constant variance.



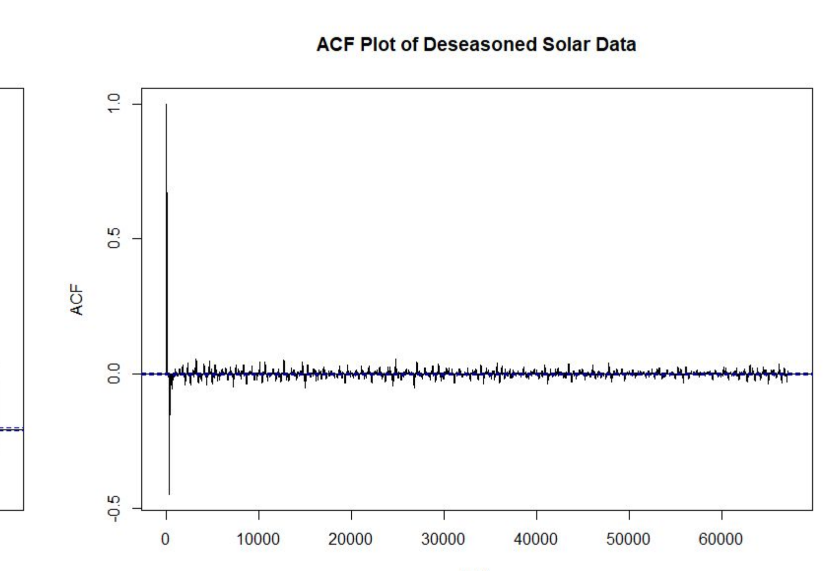
Deseasoning and detrending the data gives a mean of zero, but a non constant variance. This data has more variance than just the detrended data.

ACF Plots

These ACF plots use ¼ of the data points present in the data set.



This ACF plot of the deseasoned net demand shows that autocorrelation is heavily present within this data. This makes sense because other minor seasonalities are still present within the data to cause autocorrelation.



This ACF plot of the deseasoned solar production shows that there is less autocorrelation in this data compared to the net demand. This makes sense because removing the daily seasonality of the data should remove some of the autocorrelation.

Conclusion

This data characterization and cleansing process acts as the groundwork for a forecasting model of RIT Electricity Consumption. Understanding the data is a key step in the forecasting progress. Next steps include exponential smoothing methods, and ARIMA models. Using this data to predict peak load days can save RIT money on their electricity bill, and provide insight into their net electricity usage.