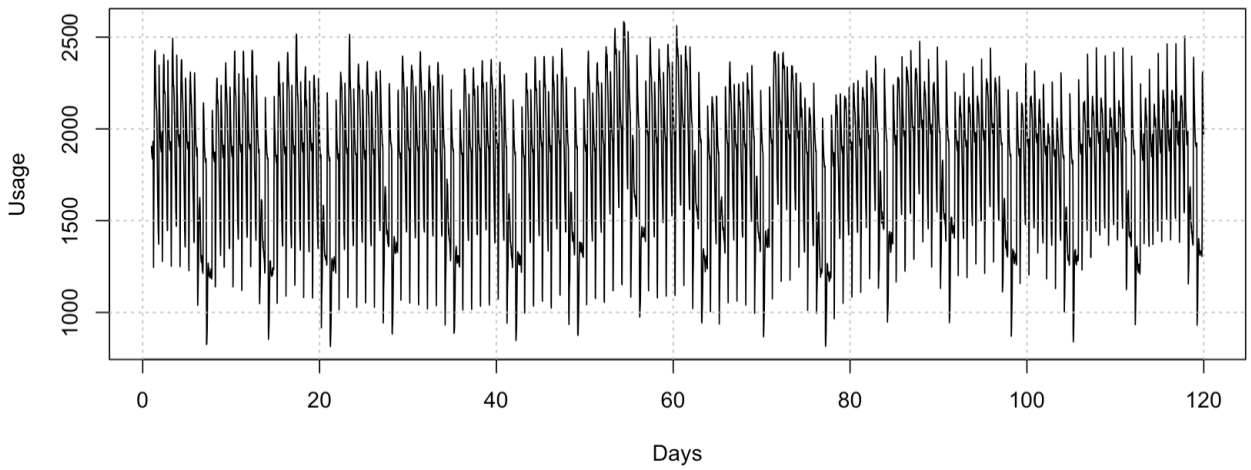
# 1. Random Forest

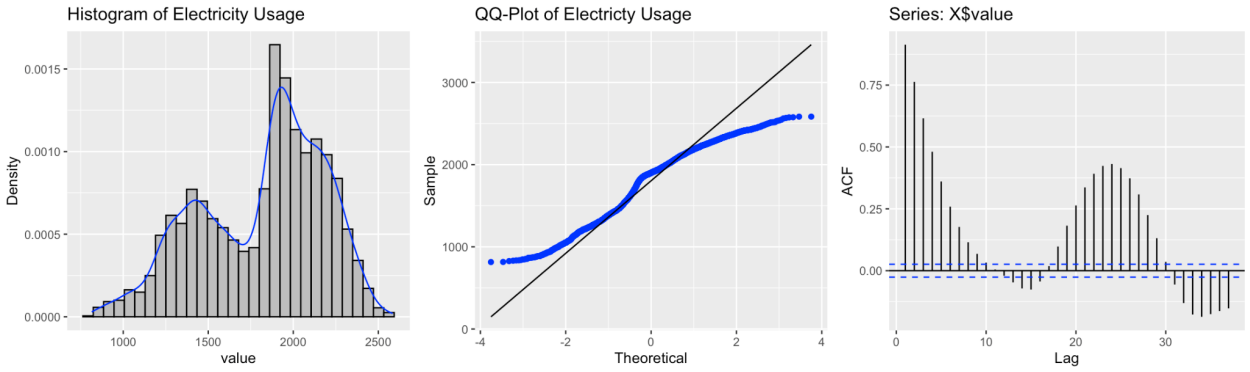
## 1.1 EDA

The DTload dataset contains 5,712 observations of electricity usage, measured every 30 minutes from May 2nd through August 28th of 2016. The data are time series: The variable of interest (value) has a variance of 133283.52, indicating it is stochastic. The variable is indexed by date, which contains sequential observations of identical periods (30 minutes). Both the variables and the index contain an identical number of unique observations: 5,712. Thus, the data are sequential observations of a stochastic variable indexed by time with constant intervals, meeting the definition of time series. Below is a time plot of the data.



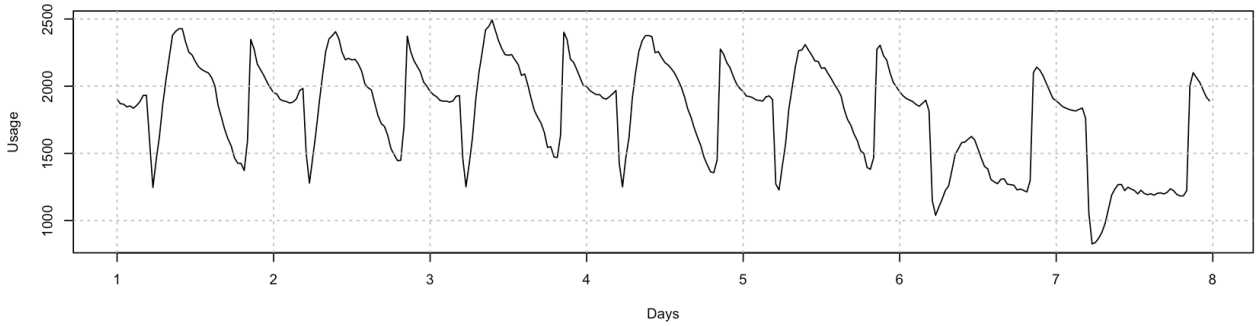

The histogram below shows that the variable value is not normally distributed: it is bimodal and not symmetric about the mean. The QQ plot in the middle shows that its tails deviate from the normal line as well. Neither the 95% confidence intervals for skew (-0.502, -0.414) nor kurtosis (-0.773, -0.623) included zero.

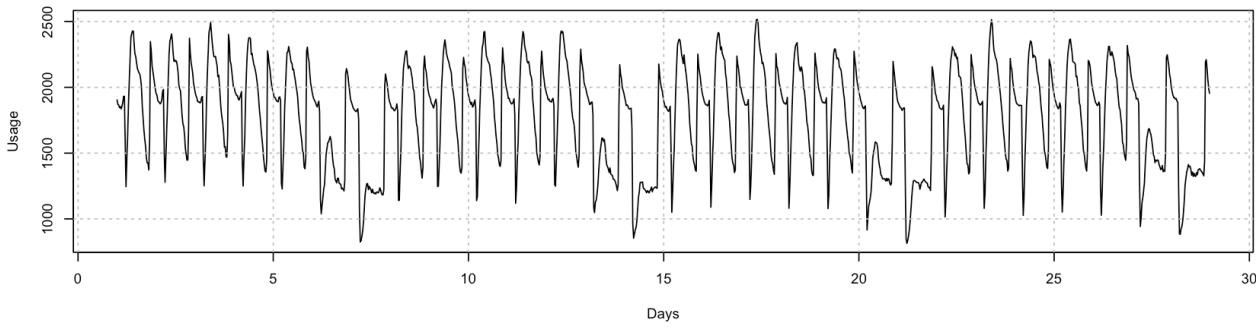
The t-test 95% confidence interval for value is (1807.074, 1826.013), thus the variable is not mean zero. Therefore, it cannot cannot be strictly stationary. However, a McLeod-Li test returned an empty set, indicating that the variable does have constant variance.
ADF (p 0.01 < 0.05) and KPSS (t .0571 < 0.146) tests both return test values beneath those required to assert linear-trend stationarity, agreeing with the notion of constant variance.

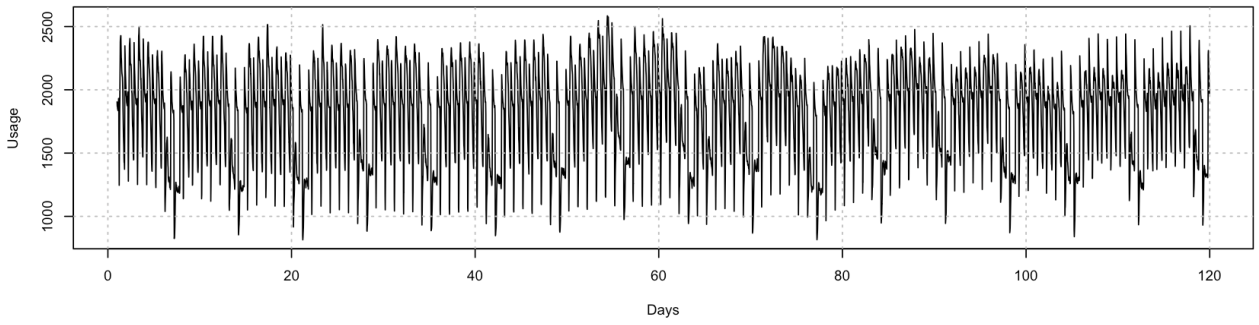


There are two seasonal patterns within the data, daily (p=48 and weekly (p=336). The time series, as well as both seasonal patterns, are plotted below. The first plot shows a clear daily pattern: usage peaks in the late morning and late evening. The drops after these two spikes have opposite convexities. In the early mornings and early evenings, usage bottoms out. There is also a weekly pattern, shown most easily in the second plot, which appears sinusoidal.

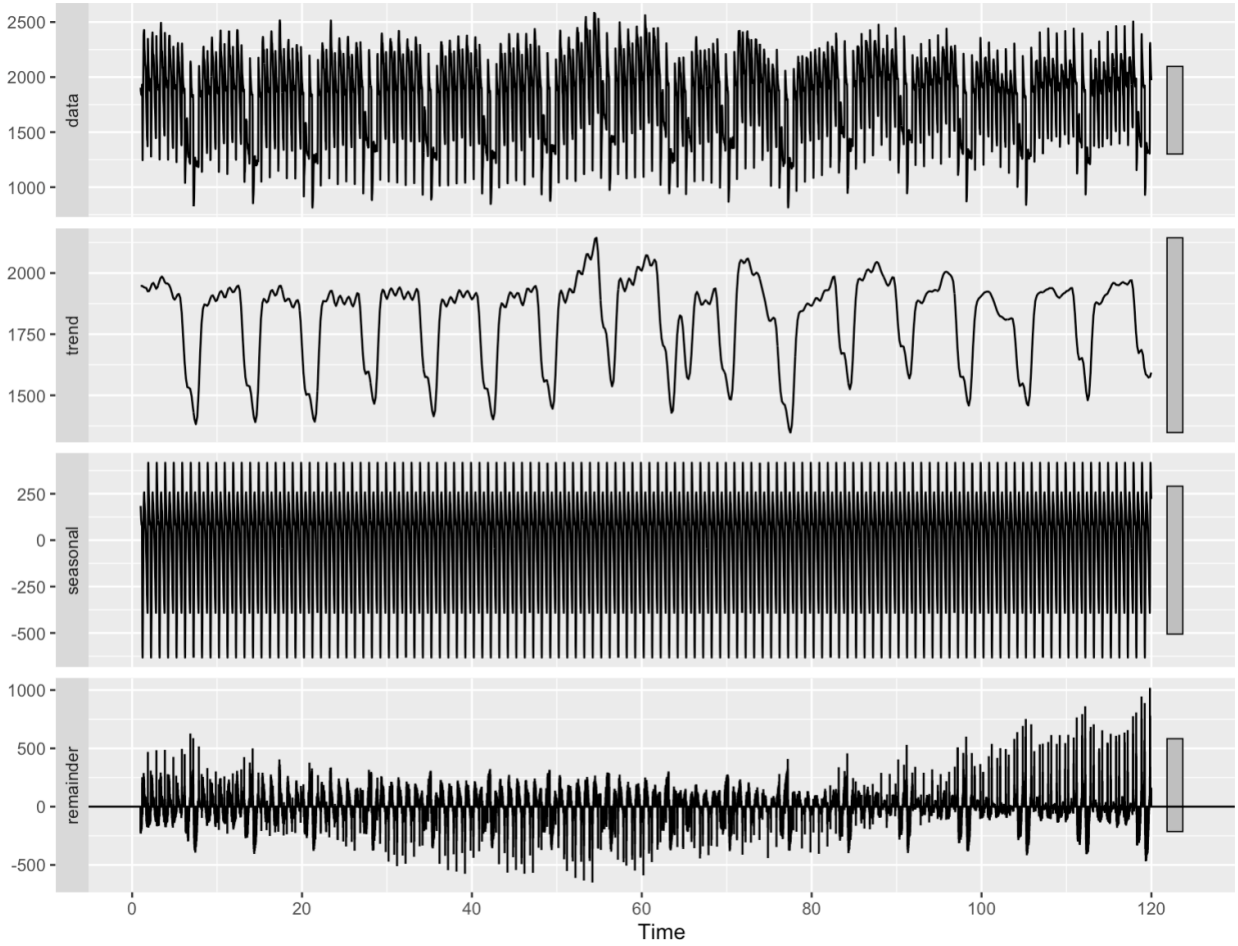
Strong seasonal patterns such as these imply high auto-correlation. The variable's ACF plot, shown on the previous page, demonstrates this to be the case. The ACF plot also displays a clear sinusoidal pattern across the lags. The period, measured from first to second peak, is 24 measurements every thirty minutes, indicating a 12-hour period. This occurs twice per day, and represents that the two spikes occur in the late morning and late evening, 12 hours apart.



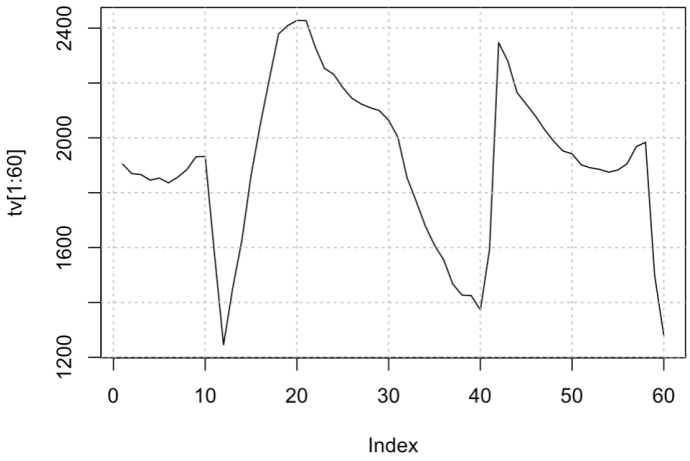



Below is a decomposition of the time series using the weekly seasonal component. The mean of the time plot stays relatively constant, yet the trend plot shows a high degree of variance, indicating that this is not very descriptive of the trend. Because this decomposition uses the weekly seasonal component, the seasonal plot displays highly regular repetitions of the weekly pattern. The residual plot does show a variety of sinusoidal patterns. The largest has a bottom around 50 and a peak around 120. Overall, this decomposition is not a great fit in terms of trend and explaining all variance, however, it does allow for seeing strong evidence of several of the seasonal components of the series.

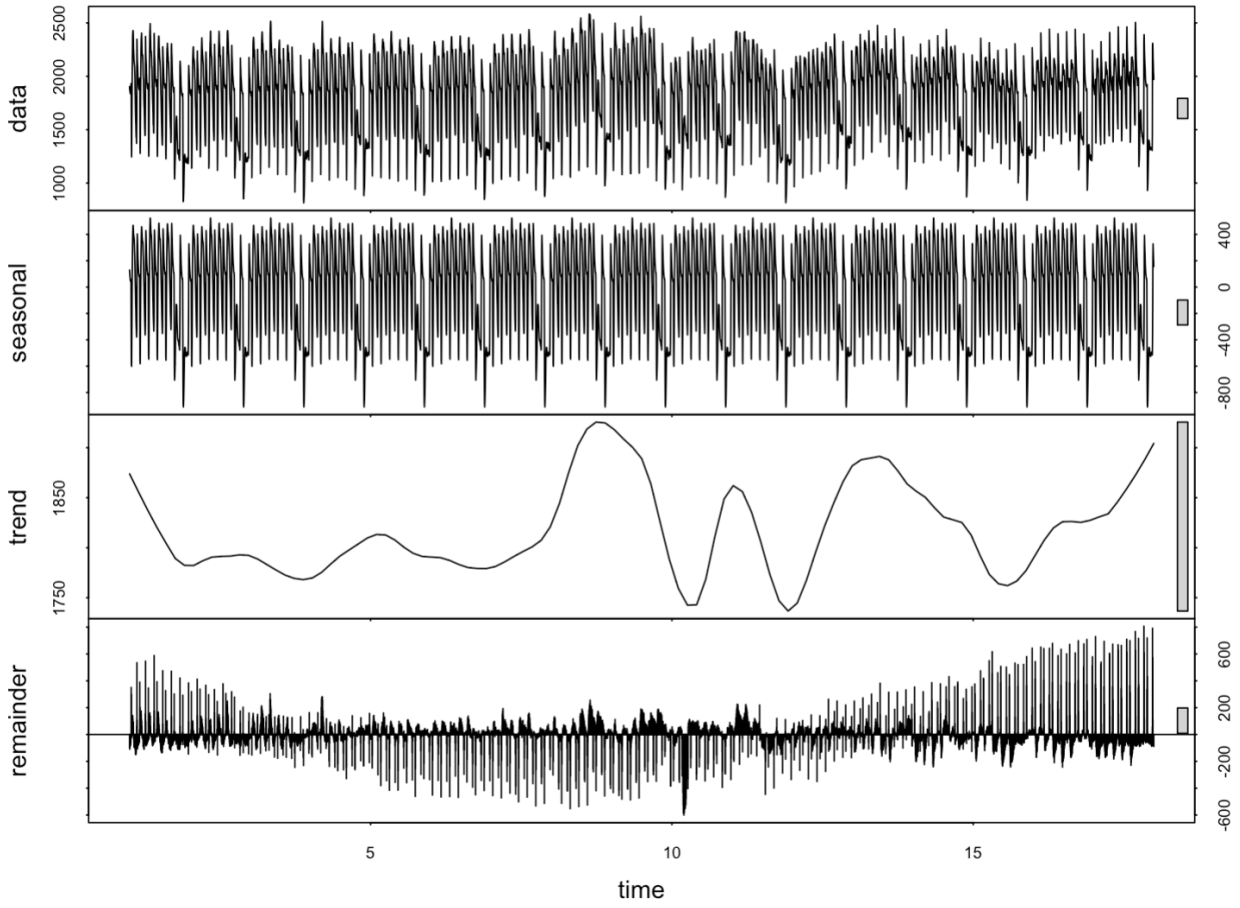


### 1.2 Fourier Analysis and Principal Components

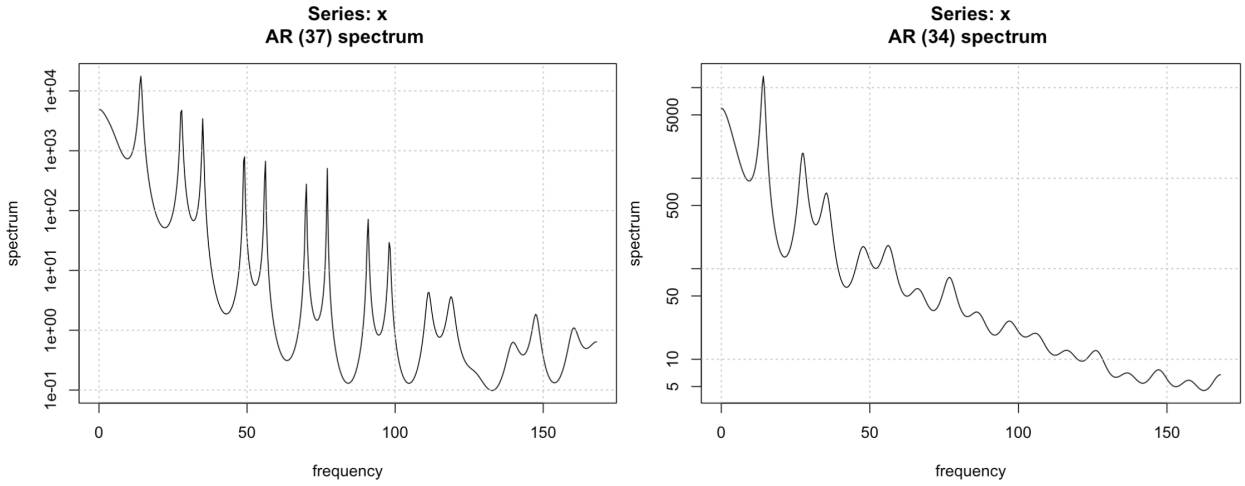
As previously described, the data have daily and weekly seasonalites. The daily pattern is shown to the right, identical to the first plot in the three on the second page.




Below is another decomposition of the data using a weely sliding window, representing the weekly seasonal component. Compared to the previous decomposition, this produced a much simpler trend line. The time and seasonal plots are virtually identical, both displaying the weekly pattern. The residuals of this decomposition more clearly depict the sinusoidal pattern that I previously described that bottoms before the middle of the time series and peaks at the end.



The two plots below are the outputs of FFT functions applied to the data with different autoregressive parameters. They both show peaks of spectral power at frequencies 12, 24, and 48. Measured every thirty minutes, these are cycles of 6, 12, and 24 hours. These are the principal components of the daily seasonal pattern.




### 1.3 Random Forest Model Creation

Model statistics of the time series forest model also show the significance of the daily and weekly seasonal components of the data. The first output below are the model's error measurements on the training set. The two outputs beneath it show the significance of the weekly and daily components (frequencies of 336, 58 respectively.) The weekly component had stronger predictive ability, but both are significant.

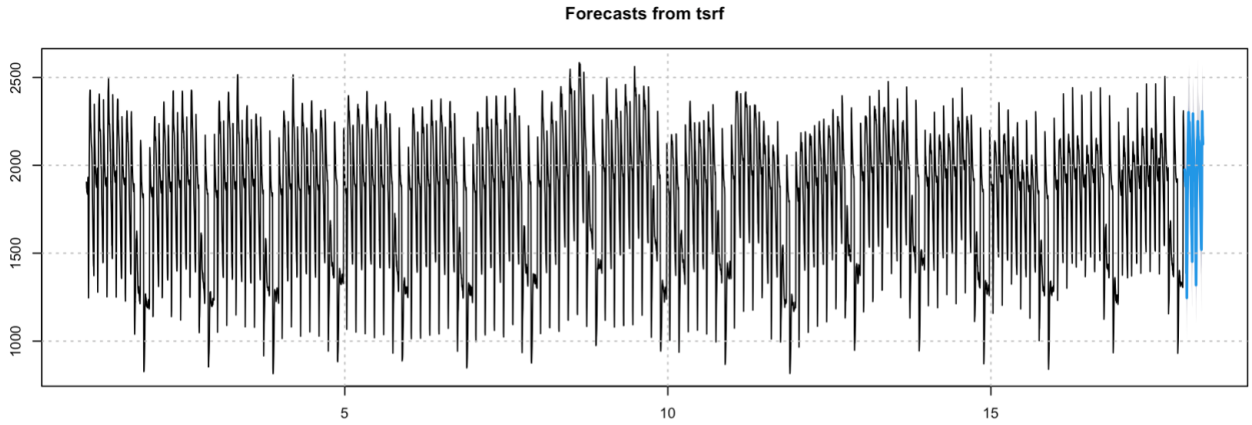
```
Error measures:
                    ME     RMSE      MAE         MPE     MAPE      MASE      ACF1
Training set 1.682635 126.1763 71.82016 -0.2163238 4.285934 0.8333613 0.5673781
```

| | Decreasing%MSE |
|---|---|
| C1.336 | 84.36774 |
| S1.336 | 54.58038 |
| C1.48 | 47.47533 |
| S1.48 | 39.98906 |
| Lag | 36.93483 |
| C1.2 | 19.96159 |

| | DecreasingNodePurity |
|---|---|
| S1.48 | 177739884.0 |
| C1.48 | 146467263.0 |
| C1.336 | 137920804.4 |
| S1.336 | 119492405.8 |
| Lag | 107235650.2 |
| C1.2 | 908181.3 |

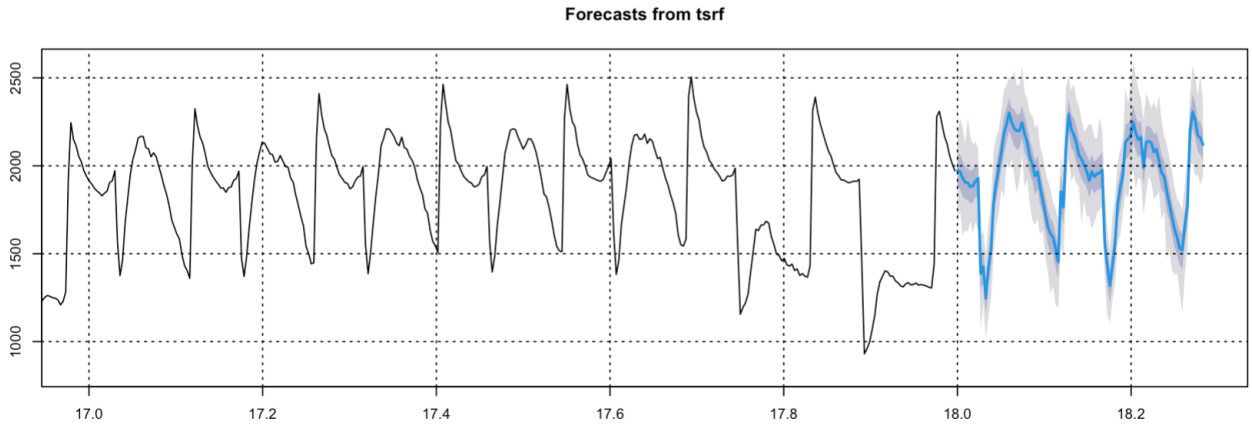
## 1.4 Forecast

The time series forest model's forecasts of electricity usage over the next two days are shown below. They look consistent with the rest of the time plot, the daily pattern is well represented.




## 1.5 Forecast Interpretation

Reducing the range of the time axis makes the details of the forecasts more discernible. The daily pattern, which contains two peaks and two valleys, is captured well by the model. Usage around the first and second peaks differ, with usage dropping more slowly after the late morning spike. The two days of forecasts maintain this pattern. The forecasts overall look more similar to the days in the beginning of the plot, with the most recent two being at a lower level. The forecast captures the average level well, but the change in level is noticeable.



## 2. Amazon DeepAR

### 2.1 EDA

The Walmart sales weekly dataset contains 1001 weekly sales values from seven different departments going from 2/5/2010 until 10/28/2012. The data are a multivariate time series: all of the seven sales variables take on values that are unique and non-constant (and thus have variance and are stochastic) and are indexed by identical time intervals (weekly).

The line plots, as well as lines of best fit, are shown below. Each time series shows a seasonal component, they are most obvious in 1, 3, and 95. Most of the time series have a slight upward trend, except for 36. None of them are mean zero, and therefore, cannot be stationary in the strict sense. The plots of 8, 13, 16, and 93 have variance that appears indistinguishable from constant, and are likely linear-trend stationary. Due to the seasonal components of the time series, they do feature auto-correlation.


### 2.2 1-Week Forecast

A forecast window of 1 week extends the date index to 11/2/2012. Creating a new dataframe of the training data with the extended date index allows for calculation of the forecast.

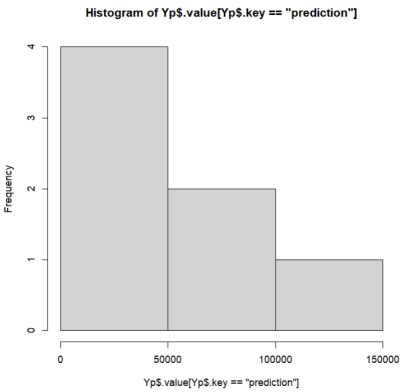
### 2.3 Model Construction and Fit

A deepAR model is fit to the data for the one week forecast using a two-week lookback period and five training epochs.

## 2.4 Calculate Forecasts

The model made the one week predictions shown below. The forecasts' histogram, shown on the right, indicates that they are not normally distributed, rather they are unimodal with high right skew. The forecasted values are in line with the levels of the timeseries.

| | .model_desc | .key | .index | .value | id |
|---|---|---|---|---|---|
| 1002 | DEEPAR | prediction | 2012-11-02 | 28509.006 | 1_1 |
| 1003 | DEEPAR | prediction | 2012-11-02 | 9878.831 | 1_3 |
| 1004 | DEEPAR | prediction | 2012-11-02 | 42399.750 | 1_8 |
| 1005 | DEEPAR | prediction | 2012-11-02 | 42346.352 | 1_13 |
| 1006 | DEEPAR | prediction | 2012-11-02 | 76327.820 | 1_38 |
| 1007 | DEEPAR | prediction | 2012-11-02 | 77036.086 | 1_93 |
| 1008 | DEEPAR | prediction | 2012-11-02 | 128269.594 | 1_95 |

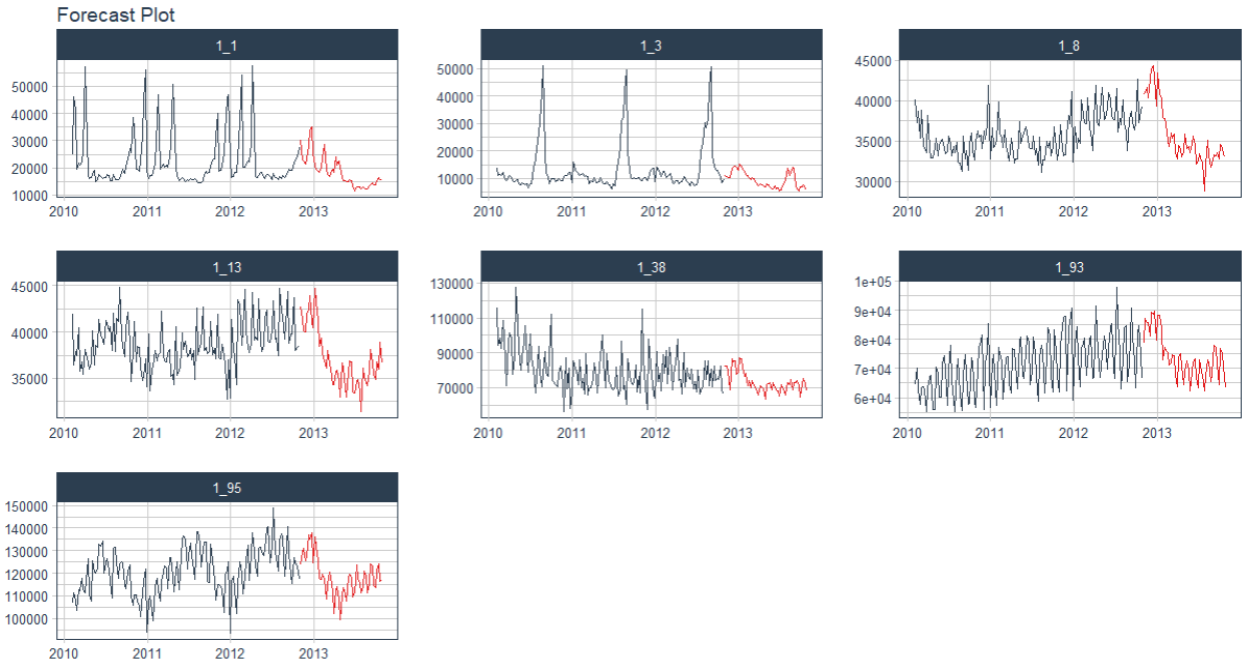



## 2.5 Plot Forecasts

Below are the plots of the one week forecasts. A single forecast does not indicate much about the pattern of variance that the model found in each time series, but the values look reasonable given the levels of the time series before prediction starts.

### 2.6a Horizon - 13 weeks

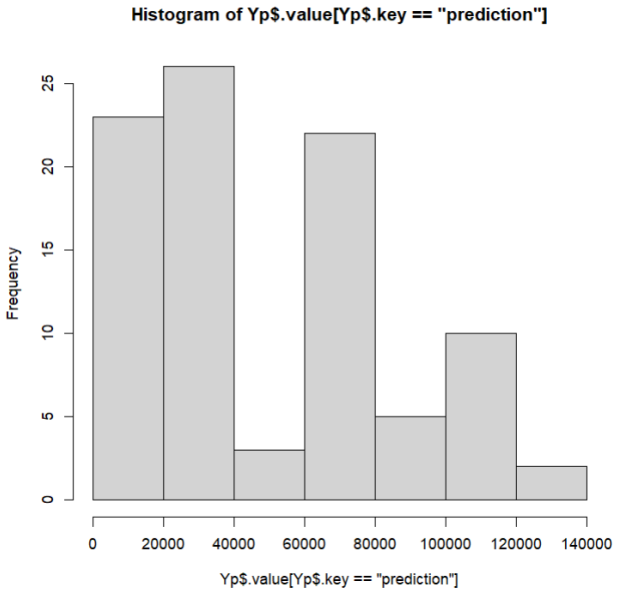
A deepAR model with a lookback period of 26 weeks and five training epochs produced the 13 week forecasts shown below. The forecasts capture the outline of seasonal components in each time series well. The distribution of spikes and their general features in the time series' seasonalities are well preserved in the forecasts. The forecast of 95 looks quite close to the variable's recent behavior. However, the magnitude of some of the spikes and overall level changes in some the forecasts are less impressive. For instance, the forecast of 1 correctly identifies the 4 spikes that occur in the seasonal pattern, but the height of these spikes is considerably lower than in the data. Another issue, most notable in 93, is that some of the forecasts appear representative of the average of the data, but do not capture recent level changes. Overall, I do think these models capture a great deal of the time series' patterns.




The output of the nonparametric test function applied to the model and its forecasts is shown below. The output indicates that at least one of the forecasts is statistically different from the others and that the variables' explanatory power increases with their ID (except for 1 and 3).

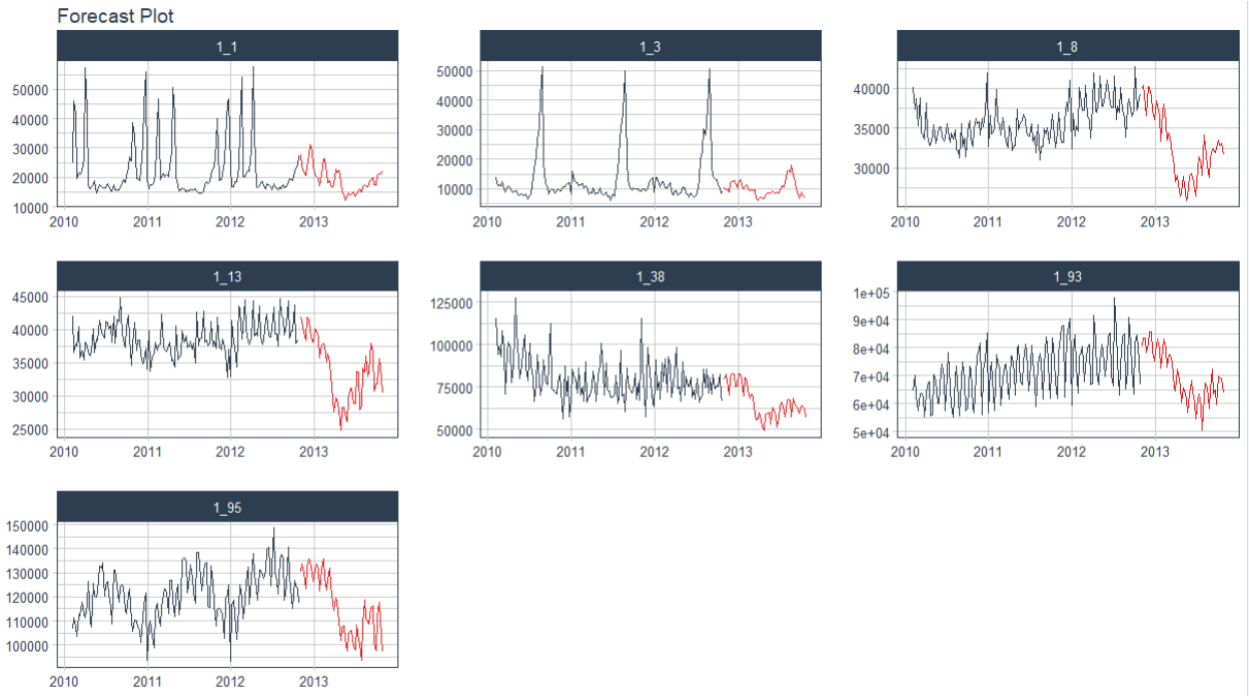
```
                                                 Test Statistic   df1 df2 P-value Permutation Test p-value
ANOVA type test p-value                                     273 6.000  84       0                       0
McKeon approx. for the Lawley Hotelling Test                273 6.000  84       0                       0
Muller approx. for the Bartlett-Nanda-Pillai Test           270 6.067  84       0                       0
Wilks Lambda                                                273 6.000  84       0                       0

$releffects
       .value
1_1  0.214290
1_3  0.071429
1_8  0.385040
1_13 0.472100
1_38 0.714710
1_93 0.713860
1_95 0.928570
```

The forecasts' histogram is shown on the right. Overall, their distribution is still highly right skewed, similar to that of the one week forecasts. Notably, the distribution of the 13 week forecasts appears almost trimodal, with low density from 5000 to 6000 and 8000 and 1000. However, the right skew from the original histogram is preserved.



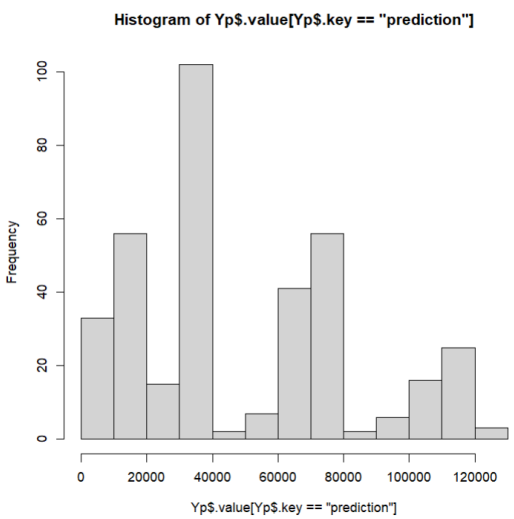

### 2.6b Horizon - 52 Weeks

A deepAR model with a lookback period of two years and using five training epochs produced the one year forecasts shown below. Similar to the 13-week forecasts, the seasonal components in 1 and 3 are well captured by the model but have too little magnitude. The rest of the forecasts are worse than their 13-week counterparts; the level of each of them precipitously falls over the first half of 2013 to eventually rise to half the average level of their respective time series. These are not patterns present in the data, and thus look like poor forecasts.



The output of the nonparametric test function applied to this model is shown below. The top tests show that at least one of the forecasts is statistically different from the others. The bottom section shows that this model agrees with the previous models in how they describe the variables' predictive ability as increasing with their IDs, except for 1 and 3.

The histogram of the 52 week forecasts, shown on the right, indicates that the overall right skew from the original histogram is maintained and that the tri-modality I suspected in the previous histogram is been maintained as well.

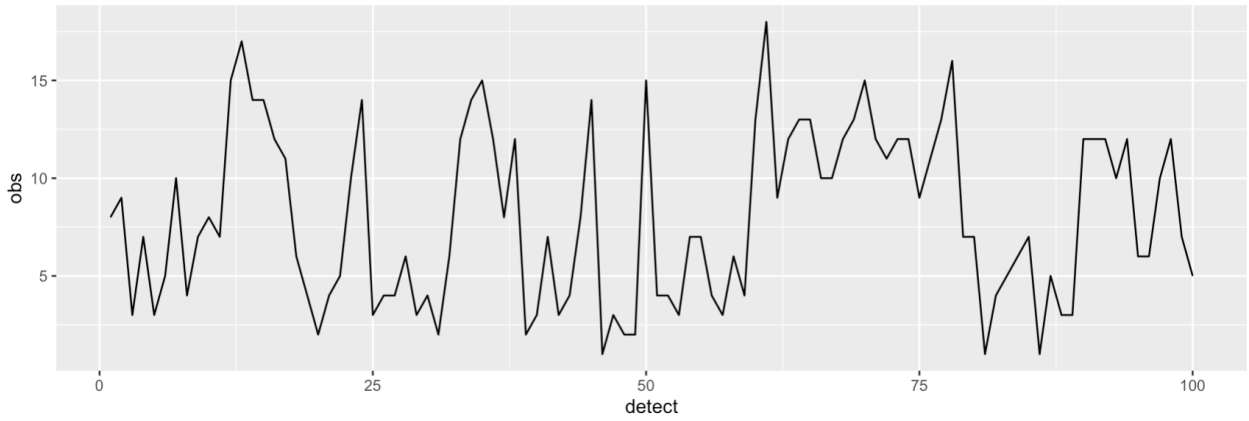



```
                                                 Test Statistic   df1 df2 P-value Permutation Test p-value
ANOVA type test p-value                                1226.892 6.000 357       0                       0
McKeon approx. for the Lawley Hotelling Test           1226.892 6.000 357       0                       0
Muller approx. for the Bartlett-Nanda-Pillai Test      1223.521 6.017 357       0                       0
Wilks Lambda                                           1226.892 6.000 357       0                       0

$releffects
       .value
1_1  0.214290
1_3  0.071429
1_8  0.385990
1_13 0.471150
1_38 0.685120
1_93 0.743450
1_95 0.928570
```

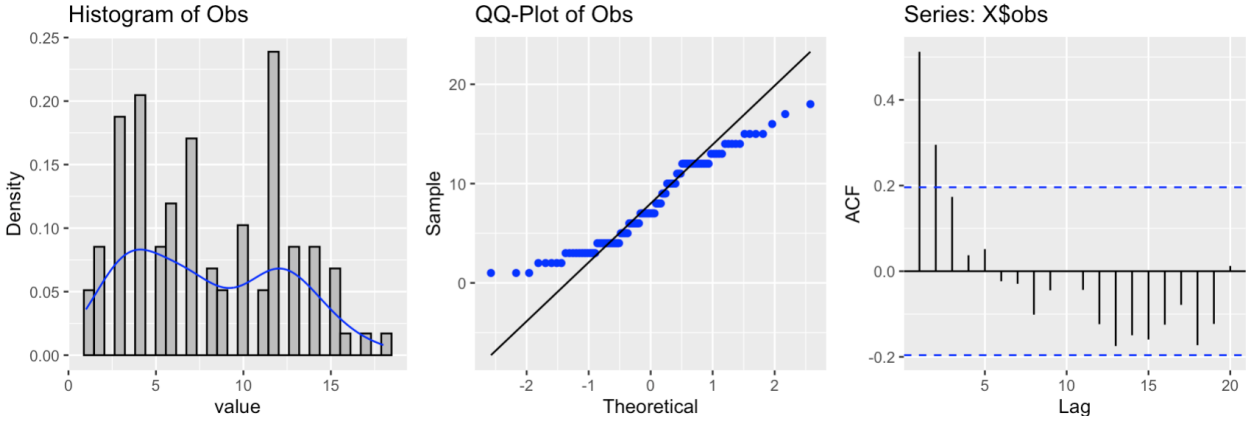
## 3. Hidden Markov Models

### 3.1 EDA

The Exo dataset contains 100 unique observations of four variables: obs, guess, state, and detect. The data are a multivariate time series. The index of the dataset is the detect variable, ranging from 1 to 100 in steps of 1. Thus, the dataset is indexed by constant intervals. The variables obs and guess are the observed and guessed dips in the light curve, and they have variances of 19.201 and 6.694 respectively, indicating they are both stochastic. The variable State is a categorical variable indicating the Exoplanet, either 1 or 2. Thus, the data are a sequential series of unique observations of variables over constant intervals, fitting the definition of time series.The variable of interest (Obs) is shown in a line plot below.



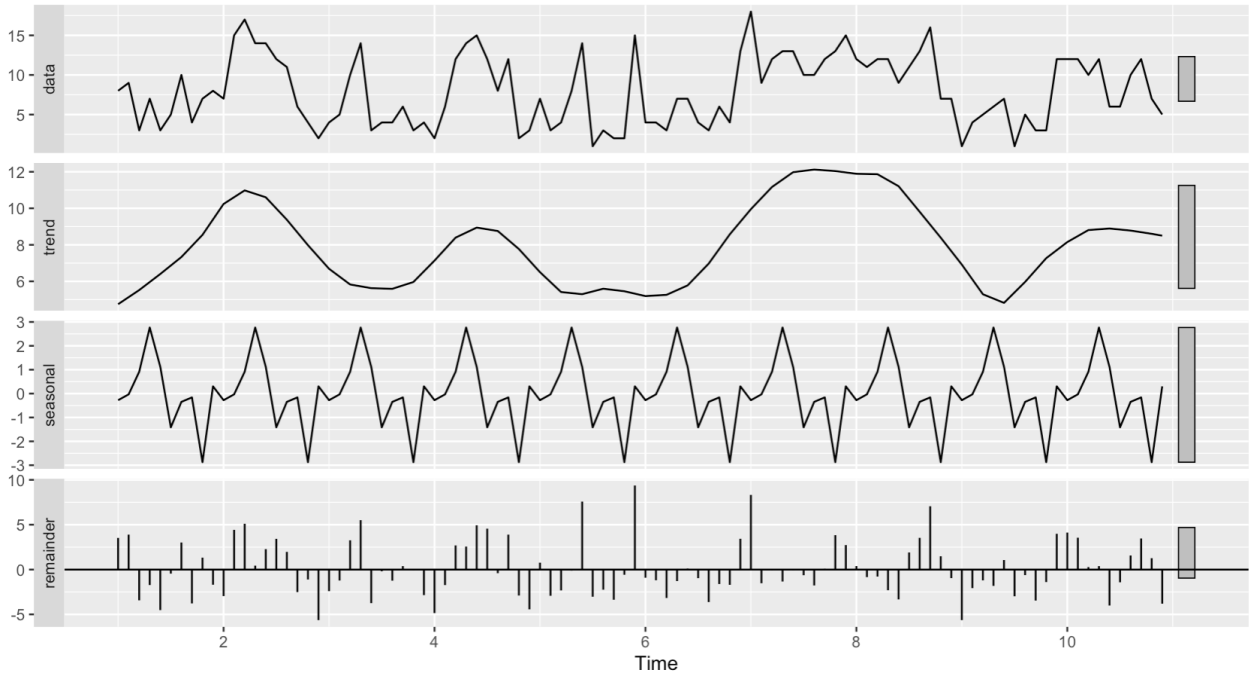

The variable Obs is not normally distributed. Its histogram below shows that it is bimodal and not symmetric about the mean. The QQ plot shows that the tails deviate from the normal line. The 95% confidence interval for skew does include zero (-0.0689, 0.555), however, it does not for excess kurtosis (-1.49, -0.842).



The variable Obs has a t-test 95% confidence interval for its mean of (7.101, 8.839). So, the variable does not have a mean of zero, implying it is not stationary in the strict sense. A McLeod-Li test returned an empty set, indicating that the variable has constant variance. This is consistent with linear-trend stationarity being suggested by a KPSS test, which each returned a test statistic beneath that required to assert stationarity ($0.066 < 0.146$).

The ACF plot on the previous page does not indicate that the variable has much auto-correlation. However, there is a clear sinusoidal pattern in the spikes that appears to have a frequency of 30 given the peak to valley in the ACF plot is across 15 lags. A seasonal component is not clearly discernible to me in the line plot, and so perhaps the strength of this pattern is weak.

Below is a decomposition of the Obs variable using a frequency of 10, which represents a year of observations (100 observations / 10 years = 10 observations / year). The trend line has a sinusoidal pattern with a frequency of about 2 years. The seasonal component captures the patterns of variance in the line plot well. The residuals are not small in range relative to the overall series, and they do exhibit a sinusoidal pattern, but I do not look at them and easily picture them within the line plot. The residuals indicate that there is some sinusoidal pattern in the data that the decomposition did not capture.



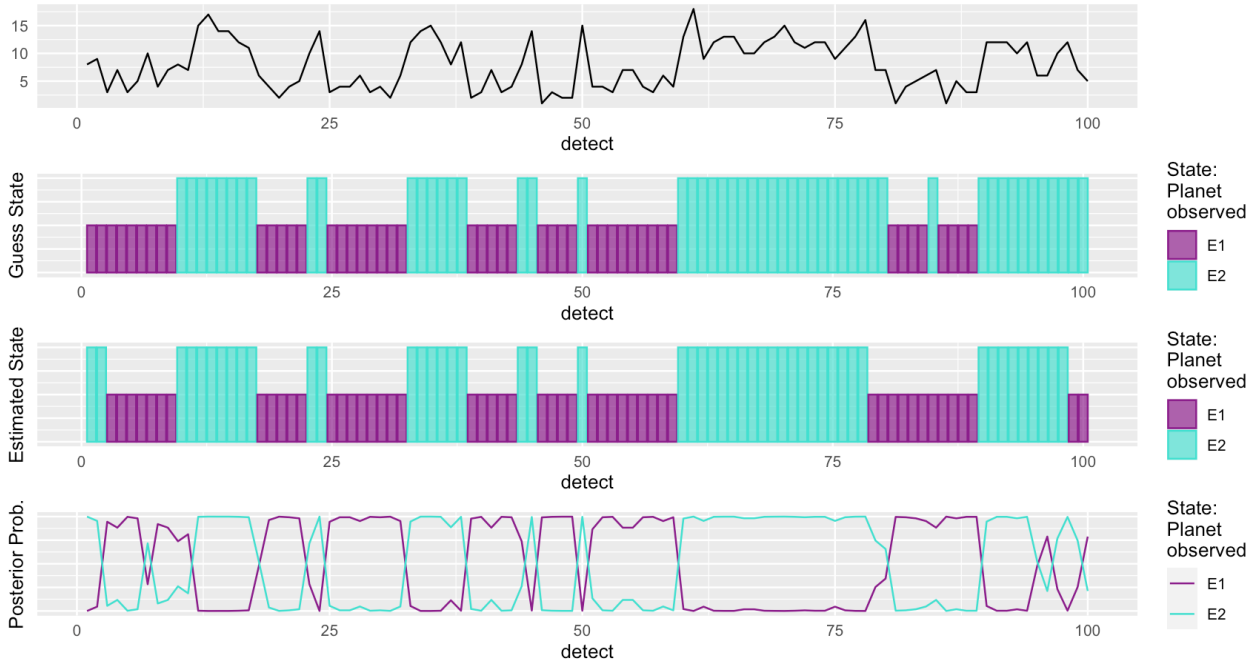
### 3.2 Model Creation

To the right is the output from fitting an HMM model to the data. The model had a specificity of 0.904 and sensitivity of 0.958, which means that it correctly predicted true positives 90.4% of the time and true negatives 95.8% of the time. The positive and negative predictive values represent the same ideas, which is why they are identical to sensitivity and specificity. The model model made two false positive and five false negative classifications. The optimal criterion for the model, which represents the parameter at which classifications are optimal, is 0.862.

| | Estimate |
|---|---|
| cutoff | 1.0000000 |
| Se | 0.9038462 |
| Sp | 0.9583333 |
| PPV | 0.9591837 |
| NPV | 0.9019608 |
| DLR.Positive | 21.6923077 |
| DLR.Negative | 0.1003344 |
| FP | 2.0000000 |
| FN | 5.0000000 |
| Optimal criterion | 0.8621795 |

### 3.3 Plots

Below are a line plot of the data, plots of the guessed and estimated classifications, and posterior probabilities based on the observed classifications. The plots of the guessed and estimated classification are quite similar, showing different classifications in just 7 out of 100 time periods. Overall, the classifications and their switches are in the same place, except for the second to last guessed E2 classification and the first estimated E2 classification. Other than that, their differences are only in precisely when classification switches are made. Both the guessed and estimated classifications match the observed classifications and posterior probabilities.



### 3.4 Analysis

Overall, the model performed quite well. As stated in my analysis of the model output, the model correctly predicted true positives and negative classifications 90.4 and 95.8% of the time. This can be seen in the PROC curve below on the left. The ROC curve indicates the point at which the model's classification parameters result in this optimally balanced rate of classifications.

The plots in the previous section indicate that the model performed better than the guesses. The model also correctly predicted the E2 state during the first two observations and the E1 state in the final two observations, which the guesses did not. The guesses also incorrectly made an E2 classification around the 87th observation, which the model did not. The only instance where the guesses were better than the model was in delaying the switch from E2 to E1 classifications around the 80th observation.