TEAM PROJECT

Part 2

Team 3
Yang Yu, Chaoyang Zheng, Meng Peng & Katherine Gimon-Valencia
March 22/2019

Introduction

This report is the second in a three-part series which systematically explores increasingly complex forecasting methods in order to accurately predict daily electricity demand for h=1 from data available on day t for the Rockland Electric Company (RECO) in Northern New Jersey. This report specifically looks at a few important changes which were made to the general approach and to the calculations of the naïve methods, followed by calculations of the appropriate smoothing methods and finally regression analysis of the explanatory variables discussed in Part 1.

The first important change in comparison to the first report is the treatment of the historical data. The decision was made that rather than to use all the data which is available as of 2005, that only the most recent data would be considered in order to better forecast h=1. Specifically, we noted that older data showed greater variance (greater presence of peaks) versus more recent years where this variance has decreased. As such, it was important to capture years such as 2014 in the training set because this year showed little variance and as such, greater resembles the recent past; the result is that the forecasting methods will be able to learn from this characteristic. As a consequence, the training set will now cover the period from 2011-2014, the validation set from 2015-2016 and the test set from 2017-2018, as shown in Figure 1. Furthermore, a decision

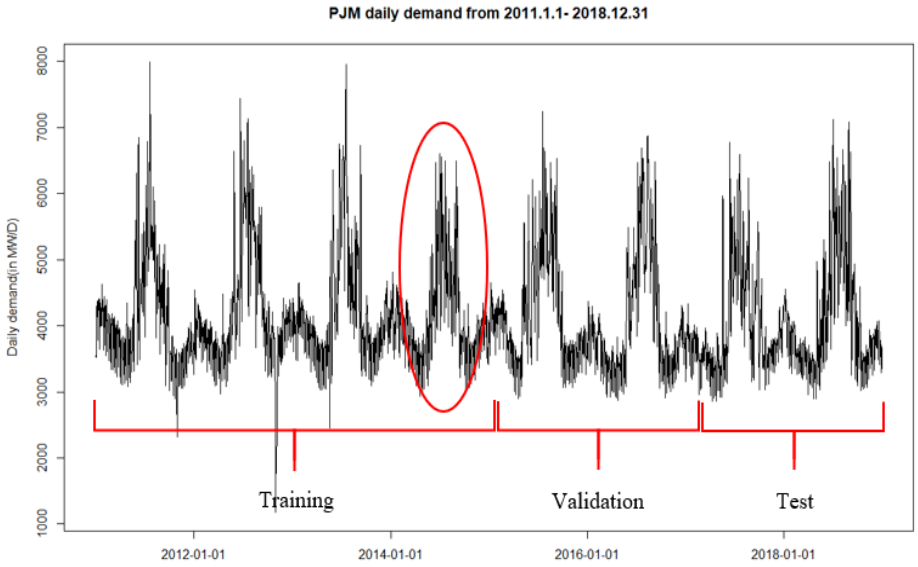

Figure 1

was made regarding how the outliers which are abnormal, that is, the dips in October 2011 and October 2012 caused by power outages due to a winter storm and Hurricane Sandy, will be treated. It was noted for example that in 2012 the demand for the entire week from October 29th to November 4th was low; this was likely due to the gradual return of the electricity and, because consumers were evacuated, this was the period of them returning home, of businesses reopening and a return to routine. Demand in the previous and following weeks were very similar on corresponding days of the week. We also noted little correlation to temperature because October is a period when consumers in this region have not yet begun heating; in other words, where the temperatures are close to the T_{ref}. As a result, the decision was taken to calculate the average demand of the same day of week from the previous and following weeks to substitute for the abnormally low week. This was done for both 2011 and 2012 as both showed similar trends and both took place in October. The last change which was made to the historical took place when we noted 7 demand values which were marked as N/A's. They were using replaced by considering the day of week and corresponding temperature values.

The next set of changes was with regards to the definition of the seasonal no-change naïve method. Previously, the same *date* had been calculated. A method has now been included to calculate the same *day* of the previous year (to better capture annual seasonality) and the same day of the previous week (to capture weekly seasonality). The results can be found in Table 1.

Table 1: Performance of 5 Naïve Methods in the Validation Dataset (2015.1.1-2016.12.31)

Method	**Bias**	**% Bias**	**Mape (%)**
No- Change	1.14	0.459	6.84
Seasonal No-Change: same date of last year	22.2	2.29	12.7
Seasonal No-Change: same day of last year	21.6	2.03	11.2

Seasonal No-Change: same day of last week	5.09	1.27	11.2
Moving Three-Day Average	1.42	0.883	9.09

All methods were recalculated to consider the new validation period. The results show that the method with the lowest MAPE, which will be the performance measure used to compare all methods and models over the course of the project, was the no-change method at 6.84%. As a result, 6.84% will be the error benchmark moving forward.

Evaluation of Smoothing Methods

The data shows that, at the minimum, any smoothing method considered must take into consideration level, trend and two seasonalities. As a result, Taylor and TBATS were considered.

Firstly, we tried state space models which are AAA, MAM, MMM. However, state space model does not applied to our data set. When we set the seasonality interval equals to 356 using the annually seasonality, the R reports that the frequency is too high.

The second method explored is double seasonality holt-winters, also known as Taylor. It was estimated with m1= 7 days and m2= 7 × 52 = 364 days to reflect the presence of the weekly and annual seasonality present in the data. We test the data for the year 2015. The Results for that DSHW model showed the bias is -126; %bias is 2.09; the mape is 8.25%. The result be found in Figure 2. The red line is predict value and the black line is observed value.

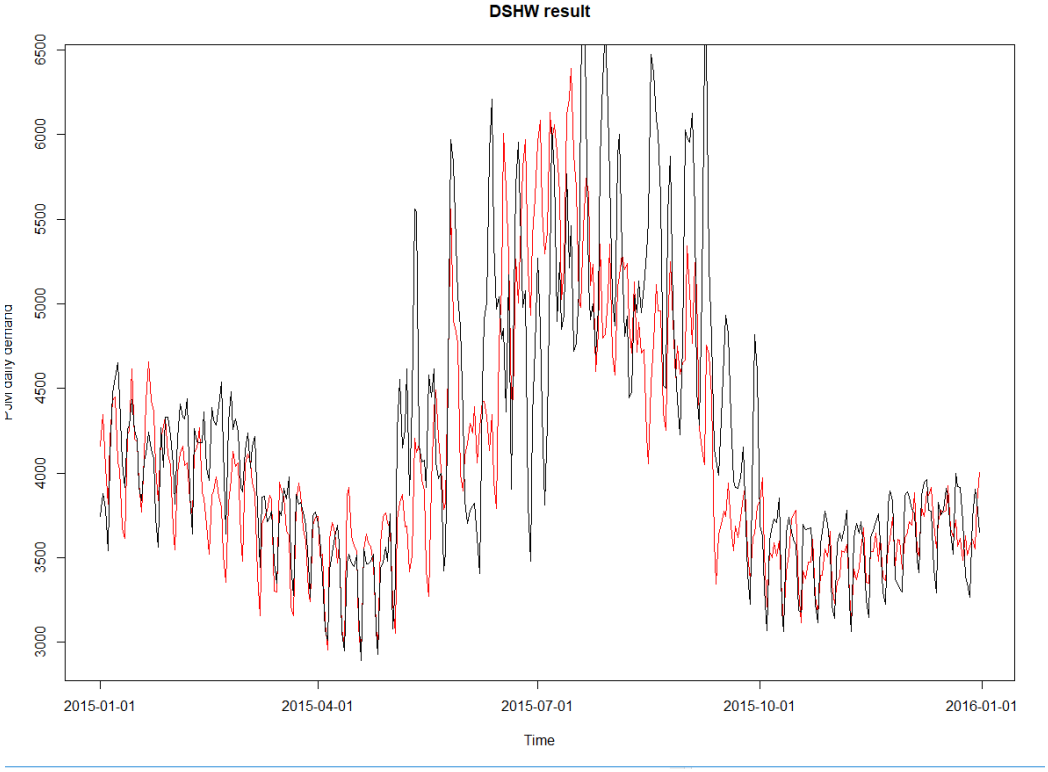

Figure 2

The last model which was tested was TBATS, with the same m values for seasonality as in the Taylor method. We run the validation set for from 2015.01.01 to 2015.01.30. We will run the whole validation data set afterwards due to the programming-crush. The Results for that DSHW model showed the bias is -152 ; %bias is -3.24; the mape is 7.58%. The result be found in Figure 3. The red line is predict value and the black line is observed value.

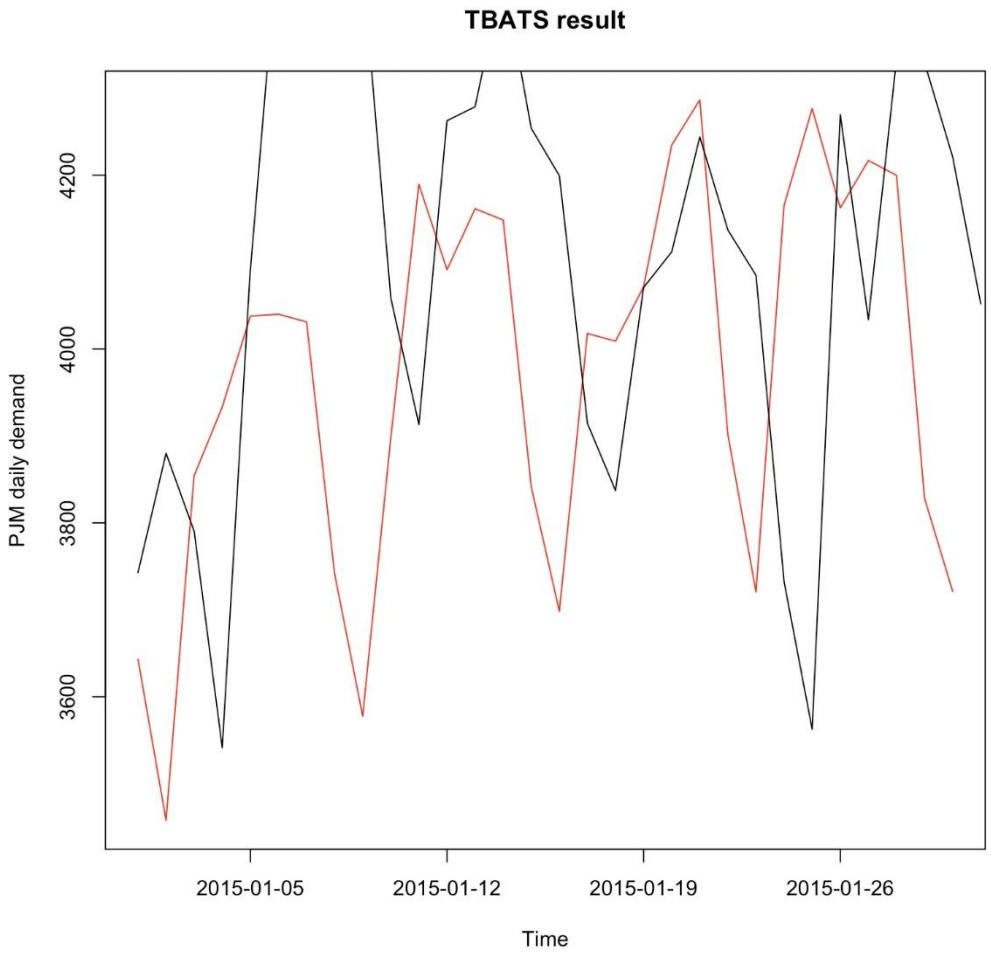

Figure 3

In comparing the TBATS and the Taylor method, we find the best to be TBATS because its mape is the smaller at 7.58%. Its performance measures in the validation set can be found in Table 2. Therefore, we see that naïve no-change method is better.

Table 2: Performance of Smoothing Methods			
Method	**Bias**	**% Bias**	**Mape (%)**
DSHW	**-126**	**2.09**	**8.25**
TBATS	**-152**	**-3.24**	**7.58**

Evaluation of Regression

In Part 1, a number of variables were discussed for consideration in the linear regression model. Before analyzing the results of the impact of these variables on the demand for electricity, a few decisions and calculations had to be made first with regards to these.

The first important decision was the determination of the T_{ref}. Figure 4 graphs the CDD with a reference temperature of 65°F. The result is a linear trend and as such, this T_{ref} was left as is. On the other hand, Figure 4 shows the HDD with the same T_{ref} of 65°F. There appears to be a flat portion at the beginning of the x-axis where consumers in our region are not yet heating; thus it appears that the 65°F is inappropriate. However, as of around 50°F we finally begin to note the emergence of a linear trend and we therefore consider this to be a better T_{ref}. The determination

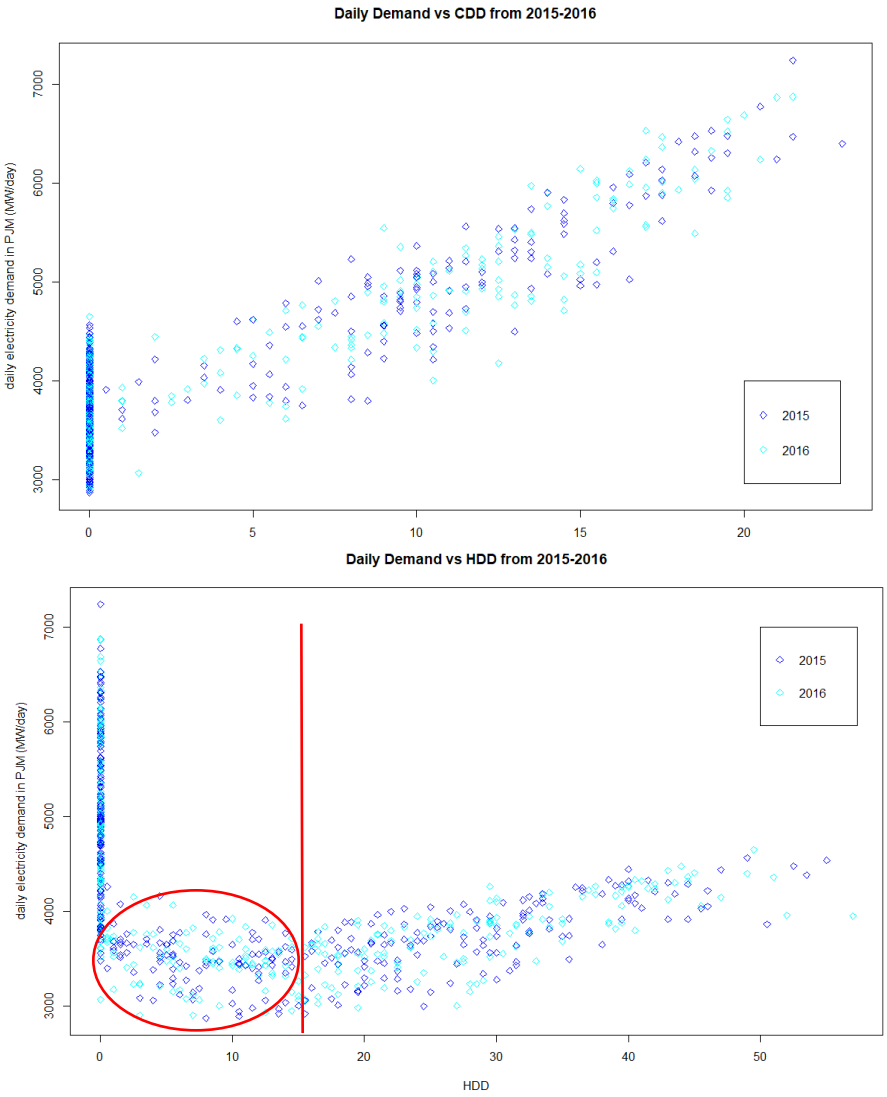

Figure 4

of the T_{ref} for both CDD and HDD at 65°F and 50°F respectively, permits us to calculate the humidity and the windchill. The humidity was calculated as the product of the relative humidity and the HDD, whereas windchill was calculated as the product of the square root of the wind speed and the CDD.

The other treatment which took place was regarding the dummy variables for the day of week. As evidenced by Figure 5, regardless of the season (even though there is some variation due to the effect of temperature) there is a distinct pattern for weekdays vs weekends and were therefore grouped as such. The dummy variable for Holidays which considers fixed, variable and days around Holidays was already discussed in Part 1.

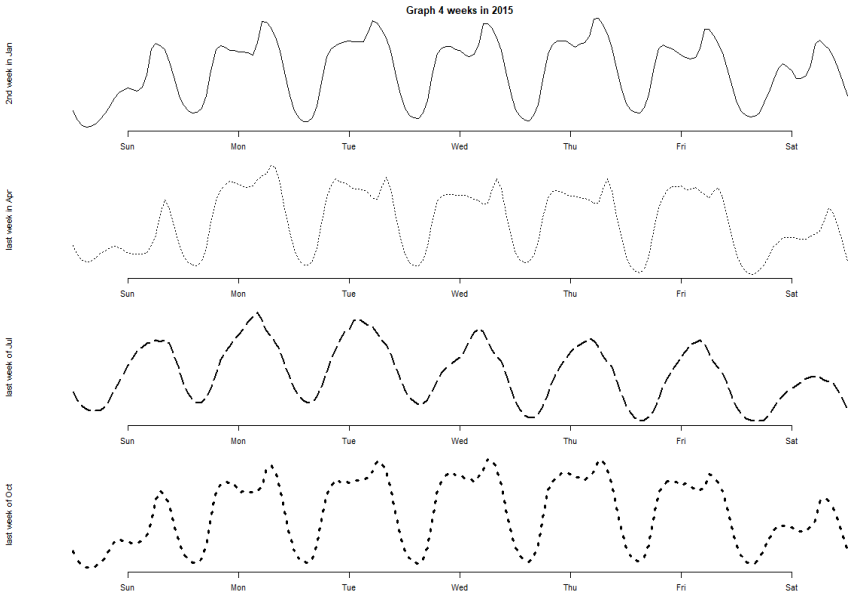

Figure 5

Calculations

To summarize, the linear model took into consideration HDDt, CDDt, lag 1 HDDt, lag 2 HDDt, lag 1 CDD, lag 2 CDD, humidity, windchill, day of week and holidays. As seen in the Figure 6 output, there are strong relationships between the dependent variable of electricity demand and all explanatory variables with the exception of windchill and lag 2 CDD (as evidenced by their large p-values). This makes sense because our data appeared to show that consumers cool more in the summer than they heat in the winter and are more sensitive to heat than cold; this result reinforces this conclusion. The R-squared for the model is 0.922 meaning there is a strong fit and a significant linear correlation.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3716.9271   10.2483  362.69  < 2e-16 ***
HDDt              18.0864    3.2804    5.51  4.2e-08 ***
CDDt             123.4448    5.6658   21.79  < 2e-16 ***
lag(HDDt, 1)       4.9798    1.9917    2.50    0.013 *
lag(HDDt, 2)       5.9303    1.4790    4.01  6.4e-05 ***
lag(CDDt, 1)      30.9007    3.1585    9.78  < 2e-16 ***
lag(CDDt, 2)       4.3684    2.2983    1.90    0.058 .
Rhumt             -0.3427    0.0808   -4.24  2.4e-05 ***
WINDCHILLt        -0.2279    0.9572   -0.24    0.812
factor(WEEKENDt)1 -552.6488   13.0971  -42.20  < 2e-16 ***
factor(HOLIDAYt)1 -218.7661   27.4620   -7.97  3.3e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 224 on 1446 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.922,   Adjusted R-squared:  0.921
F-statistic: 1.7e+03 on 10 and 1446 DF,  p-value: <2e-16
```

Figure 6

The residuals vs fitted plot in the diagnostics (Figure 7) confirms the linear relationship because the residuals are close to the 0 line. The Q-Q plot shows a largely normal distribution with a couple of outliers which are symmetrical.

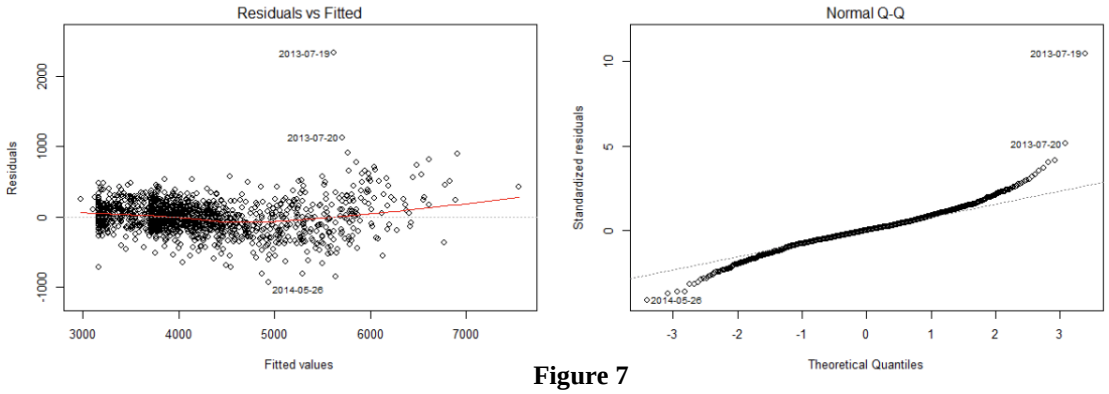

Figure 7

The next step which was taken was to run the Durbin-Watson test to check if the errors are correlated. The result of this test was a p-value of the Durbin-Watson of <2e-16, which is smaller than 0.05. As such, we reject the null hypothesis and conclude that there is autocorrelation of the errors. In order to eliminate this and avoid distorting the parameter estimates, a simple linear regression model with autoregressive errors was applied. The results of these will be discussed in the Results section.

Next, a model was tested which did not take into account the windchill (because we had seen that it was not significant) in an attempt to simplify the model. The results of this can be seen in Figure 8. The resulting R-squared is 0.922, which, similar to the previous model means that there is a strong fit and a large proportion of the total variation in the observed values are explained by the regression model.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     3716.8983   10.2443 362.83  < 2e-16 ***
HDDt              17.3895    1.4801  11.75  < 2e-16 ***
CDDt             123.4453    5.6639  21.80  < 2e-16 ***
lag(HDDt, 1)       5.0645    1.9591   2.59   0.0098 **
lag(HDDt, 2)       5.9264    1.4784   4.01  6.4e-05 ***
lag(CDDt, 1)      30.9049    3.1574   9.79  < 2e-16 ***
lag(CDDt, 2)       4.3673    2.2976   1.90   0.0575 .
Rhumt             -0.3428    0.0808  -4.24  2.3e-05 ***
factor(WEEKENDt)1 -552.6613   13.0927 -42.21  < 2e-16 ***
factor(HOLIDAYt)1 -218.5307   27.4353  -7.97  3.3e-15 ***

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 224 on 1447 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.922,   Adjusted R-squared:  0.921
F-statistic: 1.9e+03 on 9 and 1447 DF,  p-value: <2e-16
```

Figure 8

The diagnostics are plotted in Figure 9. The plots confirm that the model is linear with mean equals to 0 and with constant variance. In the first graph, we observe that the mean of the residual in the first half is

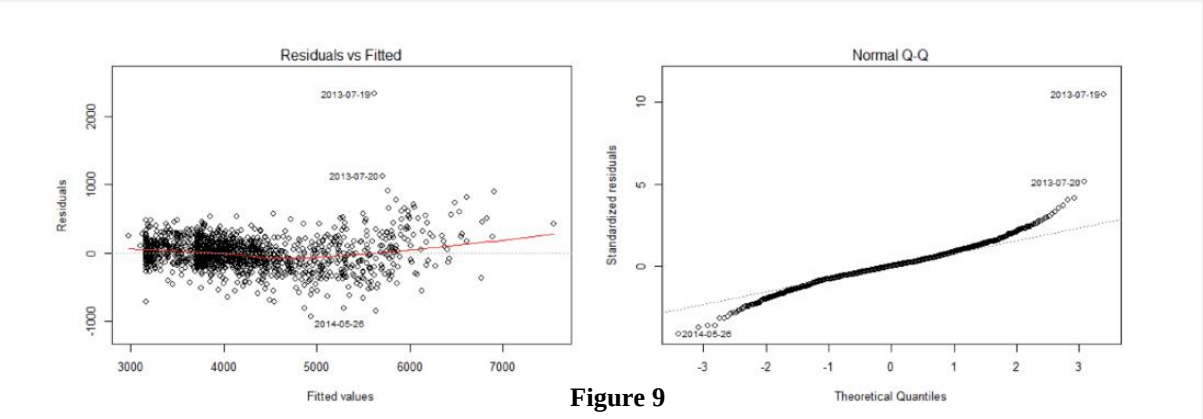

Figure 9

close to zero indicating a linear relationship. The QQ-plot again shows the residuals are largely normally-distributed however there are some outliers.

The next step was to run the Durbin Watson test which resulted in a p-value of 2e-16. Because it is smaller than 0.05, we can reject the null hypothesis and conclude that there is correlation between the errors. As a result, a second, simple linear regression model with ARMA errors was applied which will be discussed in the Results section.

Results

In Table 3, the performance of the two models with ARMA errors is compared using the AIC, AICC and BIC. The second model only yielded a value for the AIC.

Table 3: Performance of 2 Linear Regression Models with ARMA Errors

Model	AIC	AICC	BIC
All Explanatory Variables	19176	19177	19261
All Explanatory Variables Less Windchill	19225	N/A	N/A

The result is that the AIC is better for the model which included all explanatory variables. The AIC did not penalize the model for being larger. As such we determine that the model with all variables is the better of the two. However, in Figure 10, we see that the ACF of this model is still

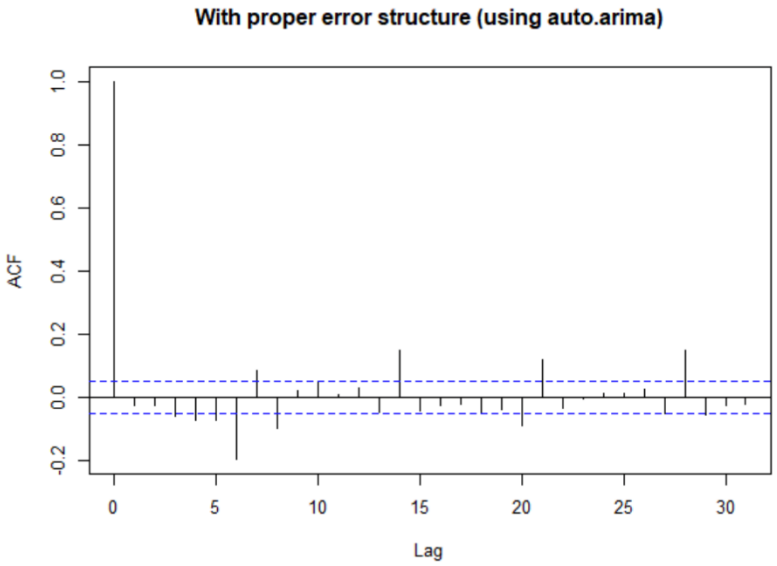

Figure 10

not stationary and there is evidence of seasonality. This will be resolved in Part 3.

The corresponding performance measures of this model can be found in Table 4. The resulting MAPE is 4.92% which is better than the no change method of 6.84%. However, because the ACF does not show white noise, this model is not appropriate as it does not meet all the assumptions. This will be adjusted in Part 3.

Table 4: Performance of Selected Linear Regression Model (2015.1.1-2016.12.31)

Method	Bias	% Bias	Mape (%)
Linear Regression Model with All Explanatory Variables	109	3.14	4.92