Forecasting Gas Prices

STAT 443: Group YAbsent

April 6, 2015

Nathaniel Horvath	20384644
Dong Kyum Kim	20359709
Wenda Xu	20419302
Herman Li	20456181
Adil Virani	20398500

Responsibilities

Nathaniel Horvath

- Performed data analysis using R (Regression, Smoothing, Box-Jenkins)
- Designed PowerPoint slides
- Compiled and formatted final report

Dong Kyum Kim

- Designed PowerPoint slides
- Wrote Smoothing Analysis section of report

Wenda Xu

- Performed data analysis using R (Box-Jenkins)
- Designed PowerPoint slides
- Wrote Introduction and Conclusion sections of report

Herman Li

- Performed data analysis using R (Regression, Smoothing, Box-Jenkins)
- Wrote Box-Jenkins section of report

Adil Virani

- Designed PowerPoint slides
- Wrote the Regression section of the report

Contents

Table of Figures

Introduction and Motivation

Ontario consumers love to complain about the price of gas, as it impacts nearly everyone, whether it is those who commute long distances to work every day, or indirectly, those who take public transportation, for instance. A lot of people wonder how gas prices are decided, and it appears there are numerous factors that go into making such a decision. Our team poses the question on whether or not we can predict the price of gas using just its history. Specifically, our goal for this project is to use the historic gasoline price in a region in Ontario to build a model to forecast the future price for the interest of consumers.

Data

Each data point represents the average price per litre of Ontario Regular Unleaded gasoline in cents. Figure 1 provides a plot of this time series.

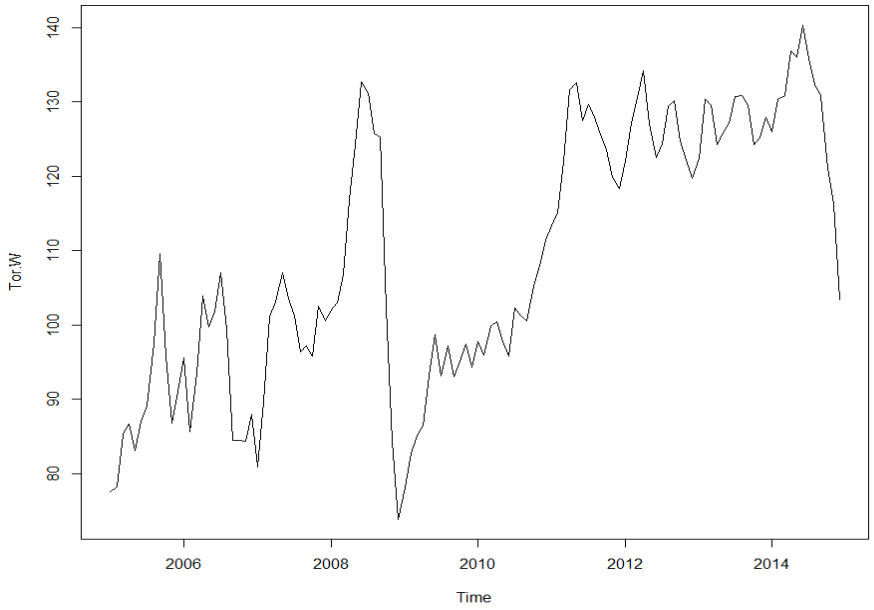


Figure 1: Plot of Gas Pump Price in Toronto West, 2005-2014

Our data is retrieved from the Ontario Ministry of Energy, and it is a set of monthly data from 2005 to 2014. It is sampled from the region of West Toronto because we believe that it is the most appropriate candidate to represent the whole of Ontario. We recognize that other locations in Ontario cities may have higher average prices that rural Ontario pumps, but we assume that the general trends we find can still be applied taking this into account.

As observed from the plot, there is an increasing trend of prices as time passes, albeit hidden behind much variance in price. Furthermore, there may be seasonal components that need to be verified during the analysis. There's also huge variation with price increase and decrease during certain periods of time. For example, the drop in 2008/2009 matches with the beginning of the recession, and the 2014 drop matches with the recent weakening of the Canadian dollar compared to the US dollar. These points are left in our data set, and not removed as outliers, as they pay homage to the fact that gas prices are volatile, and our model might not be very useful if we do not try to take into account for any sporadic spikes or dips in price.

Lastly, we note that we break up our data into training and testing sets. We use our training set to build our models, and this includes the data from 2008 to 2013 inclusive. We let the 12 data points from 2014 represent our testing set, in which we will compare our predicted model values to actual values to test model performance. Once we have chosen our best models, we will refit them to the entire 2005-2014 dataset and forecast gas prices for 2015.

Regression Analysis

As one of the methods to fit data, we perform regression analysis using several possible models and figure out the best linear model we could use to model our data.We first try the model with just the seasonality components. This involves assigning each month to an indicator variable (using 11 indicators and the intercept to represent the 12^{th} month) and then fitting the model just based on those. We can see the fit in Figure 2 in which we can see that it just takes into account the different seasons and ignores any other trend fluctuations.

Looking at the residual diagnostics in Figure 3, we can see that it suggests that the residuals are not random in nature and there might be some runs, thus violating the independence assumption. Also, the auto-correlation function (ACF) shows significant correlation between residuals, although the partial auto-correlation function (PACF) seems to indicate the opposite (which is what we expect). In the QQ plot, the quantiles do not match the QQ line at all, indicating a violation of the regression normality assumption. As we found trend in the residuals, our next model would try to eliminate this problem.

The second model will attempt to take care of the trend component, so it will have the seasonal indicator variables and also the linear time variable t. We can see the fit on this model in Figure 4, where it is an improvement but it still doesn't match with much of the variation in the data. Going over the residual diagnostics observed in Figure 5, we can see that the residuals seem to follow a quadratic relationship now. Furthermore, the ACF still demonstrates very significant correlation (while the PACF does not, as expected). The QQ plot indicates that residuals are still likely not normally distributed, as we still observe a left heavy tail.

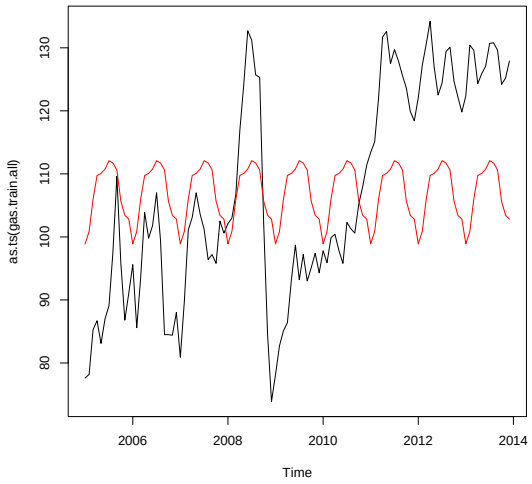


Figure 2: Fitting Seasonal Indicator Model to Training Set

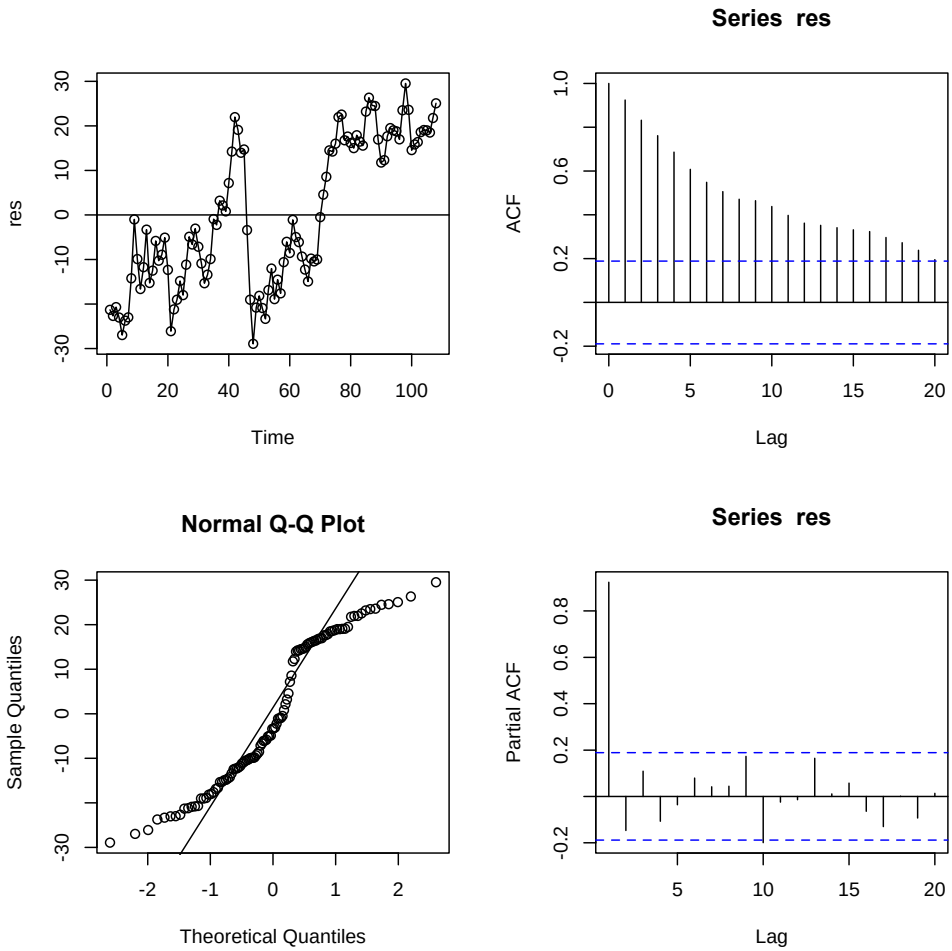


Figure 3: Residual Plots for Seasonal Indicator Model (Figure 2)

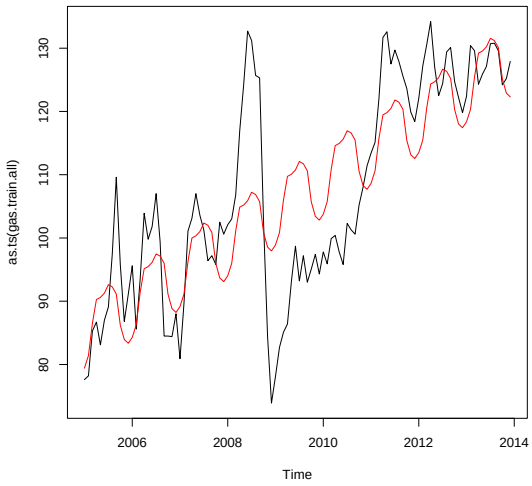


Figure 4: Fitted Linear Model, Time and Seasonal Indicators

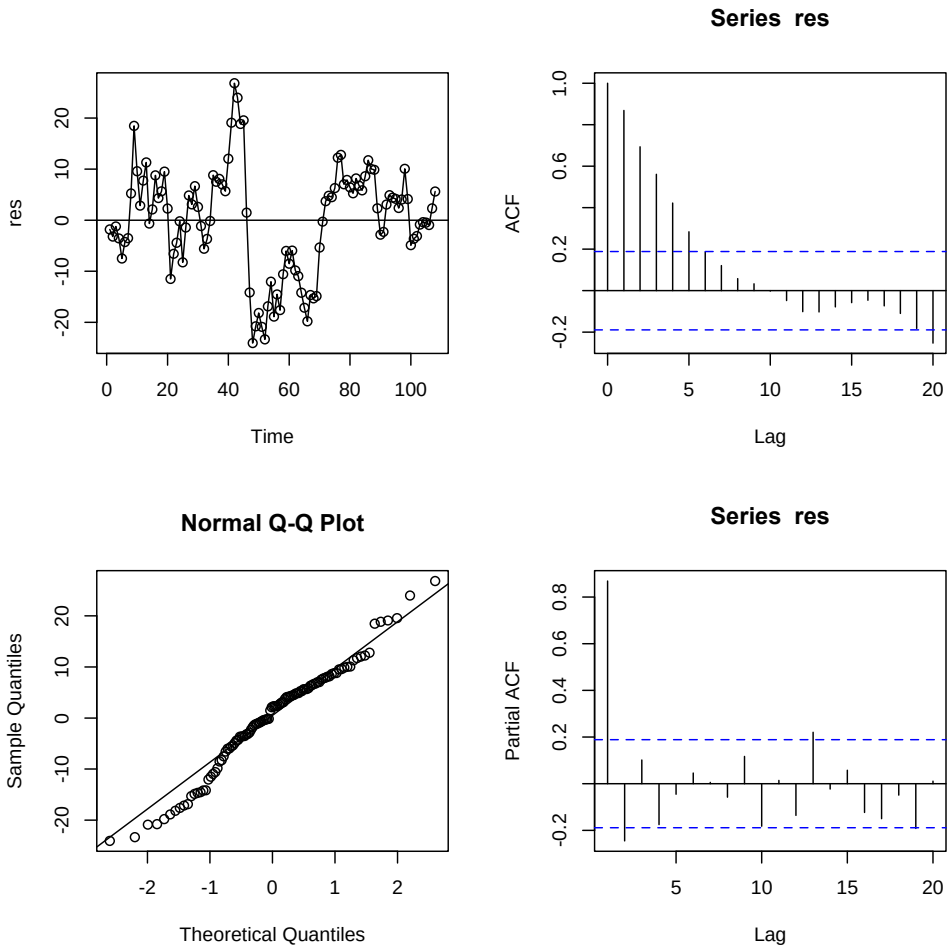


Figure 5: Residual Plots for Figure 4 Model

The third model is the same as our second model except that it also contains a quadratic time component t^2, so it gives a little better fit than the previous model as we can in Figure 6. Observing the residual diagnostics in Figure 7, the residuals seem a little better but still we see some runs, indicating potential non-randomness. The ACF still shows significant correlation. Normality also seems more likely now, as our QQ-line is matching up nicely with the quantiles.

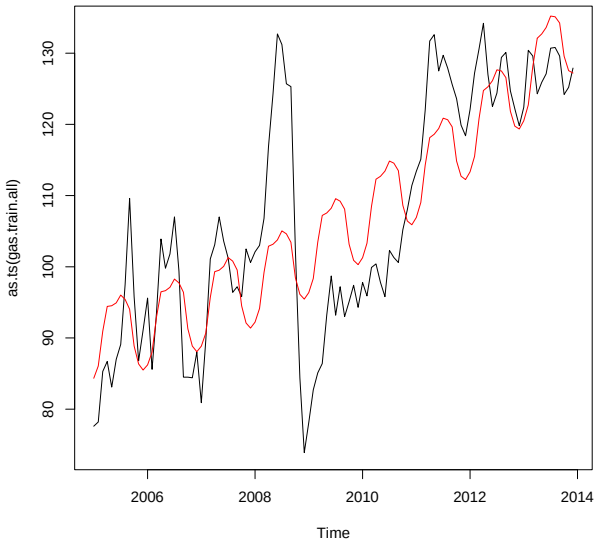


Figure 6: Fitted Model (Seasonal Indicators, Linear, Quadratic Components)

The best model we have fit is the one where we have also introduced the sinusoidal component, $\sin(t)$, to the seasonal indicators, linear and quadratic terms. This fit can be seen in Figure 8, where it seems to be a better fit than any other model so far, as it takes care of most of the variation when ignoring the events in 2008. We can also have a look at the residual analysis in Figure 9 where we can see that residuals against time still do not appear to be random about zero. Moreover, the ACF still shows significant correlation and also we can see that the normality is worse off than the previous model examined, as our upper tail drifts away from the QQ line.

To investigate whether or not the sinusoidal component or the indicator component of our previous model is best, we will also fit a model with linear, quadratic, and sinusoidal components, which gives us the fit as observed in Figure 10. Examining the residual plots in Figure 11, the residuals do not seem random at all, and we see no improvement in the poor normality of our previous model. Furthermore the ACF still shows significant correlation.

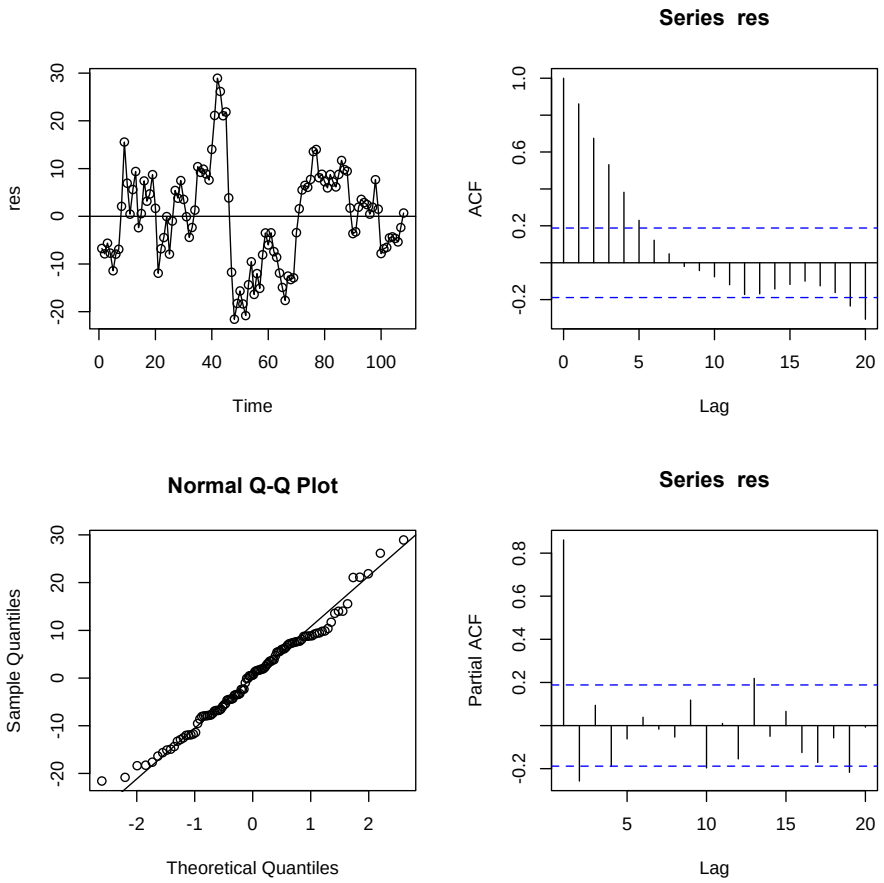


Figure 7: Residual Plots for Figure 6 Model

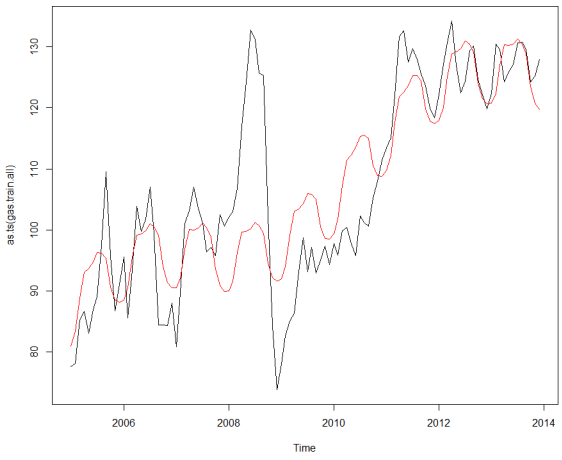


Figure 8: Fitted Model (Seasonal Indicators, Linear, Quadratic, Sin Components)

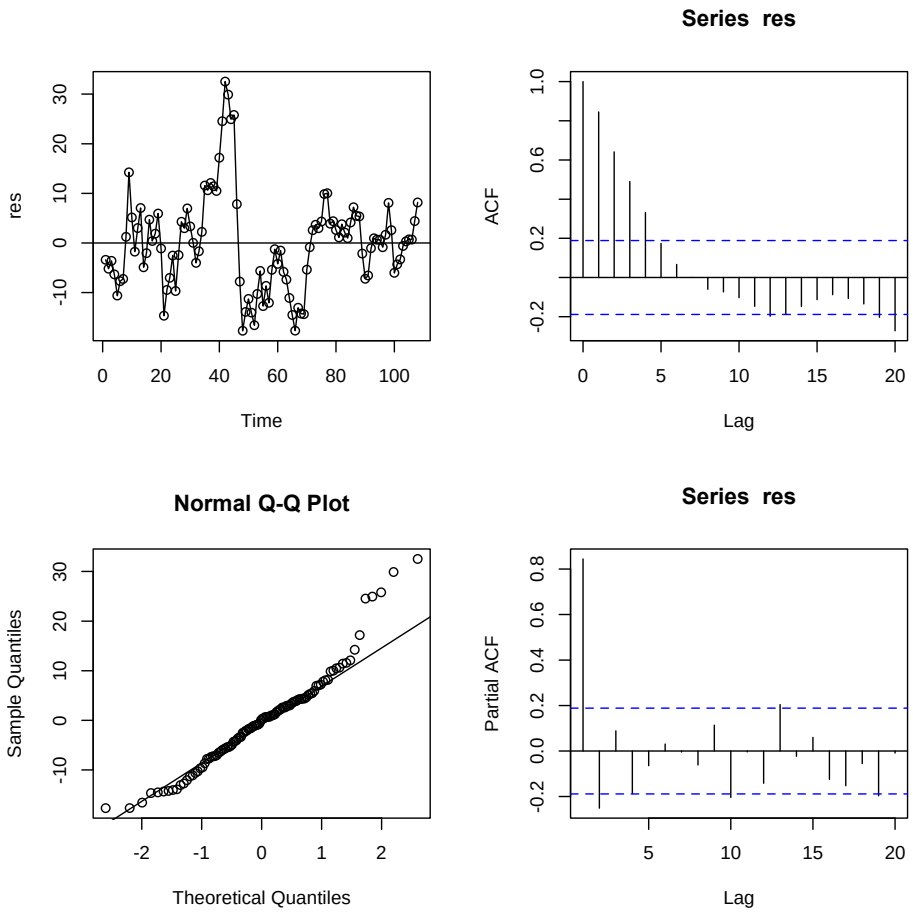


Figure 9: Residual Plots for Figure 8 Model

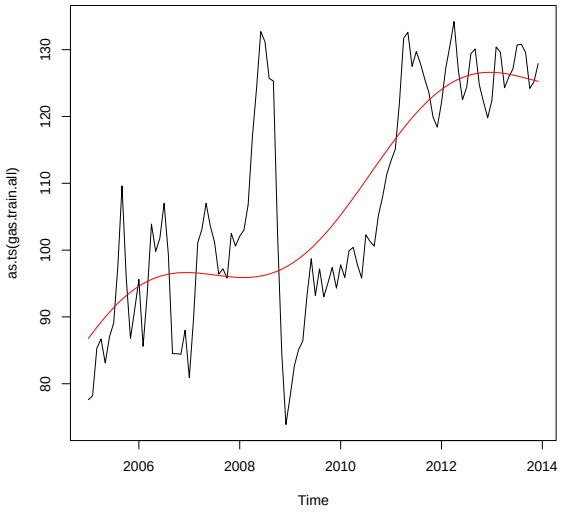


Figure 10: Fitted Model (Linear, Quadratic Sinusoidal components)

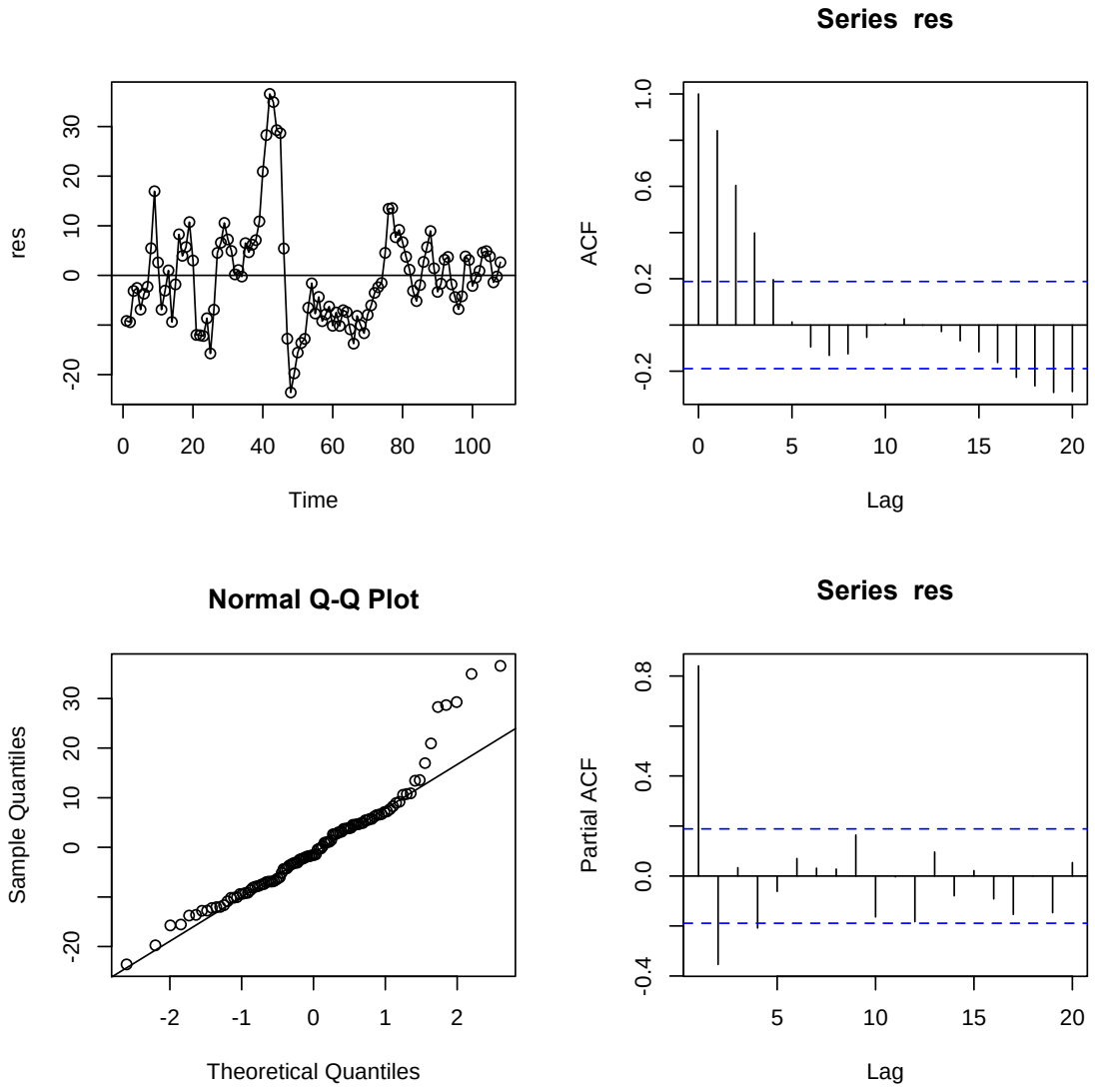


Figure 11: Residual Plots for Figure 10 Model

Now that we have five regression models analyzed, we would like to determine the best out of these five, and use that best model for forecasting. In order to determine the "best" model, we will use the following parameters to test model fit: adjusted R^2, standard error, AIC, and BIC. The values for each model are taken from the R output and summarized in Table 1. The best models in each category have their values highlighted in yellow, while the second-best models in that category have their value highlighted in green. We can see that the fourth and fifth models we analyzed perform the best here. Between these two, we decide to move forward with our fourth model that combines all of the elements tested above. Although the residuals demonstrate that the model is far from perfect, it is just a better fit to our data set.

Table 1: Comparison of regression models

	Seasonal Indicators (Figure 2)	Seasonal Indicators with Linear Component (Figure 4)	Seasonal Indicators with Linear, Quadratic Component (Figure 6)	Seasonal Indicators with Linear, Quadratic, Sinusoidal Component (Figure 8)	Linear, Quadratic, and Sinusoidal Component (Figure 10)
Adjusted R^2	-0.0399	0.5753	0.5917	0.6268	0.6013
Standard error	17.2676	11.0356	10.8208	10.3445	10.6917
AIC	935.1177	839.2828	835.8946	827.0153	824.2206
BIC	969.9855	876.8326	876.1265	869.9294	837.6313

Also, to test the effectiveness of our model, we predict the values for 2014 and compare to our testing set, as seen in Figure 12, where our testing set is in black, our model prediction is in red, and the prediction intervals are in blue. We observe that our actual value falls above the fitted value for the first half of 2014, and then falls below the fitted value for the latter half of the year. In fact, the actual value surpasses the prediction interval by the end of the year.

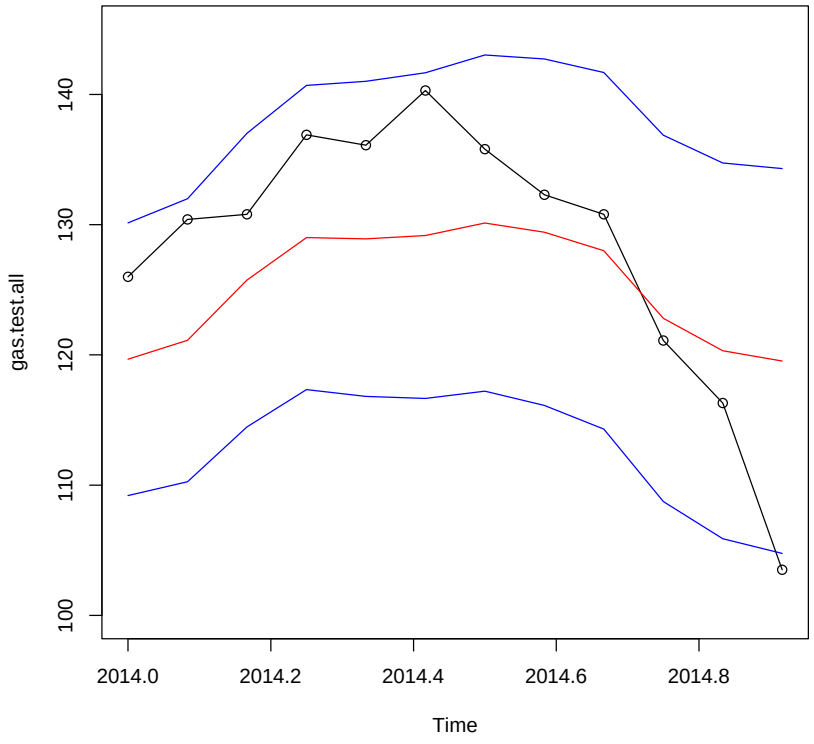


Figure 12: Best Regression Model Forecasting Testing Set with Prediction Intervals

Smoothing Analysis

We next try various smoothing methods to see if we can find a good model that fits our data. The best model we find in this section will be used to forecast 2015 at the end of this report.

We first note that we can immediately eliminate model types to test simply from what we know of our data and what we are trying to accomplish. We first note that a slowly drifting mean model is essentially useless in our case, since our data has too much variability in the mean as seen from the plot, as seen in Figure 1. Furthermore, although an MA filter is nice to use to see if we can identify any general trend in the data, it is not useful for forecasting, as the values computed for the filter involve both the past and the future. However, we perform the analysis anyways to see what trend we can spot, as shown in Figure 13. This analysis confirms that our data is generally increasing, but this trend is complicated by peaks and valleys that may be affected by seasonality.

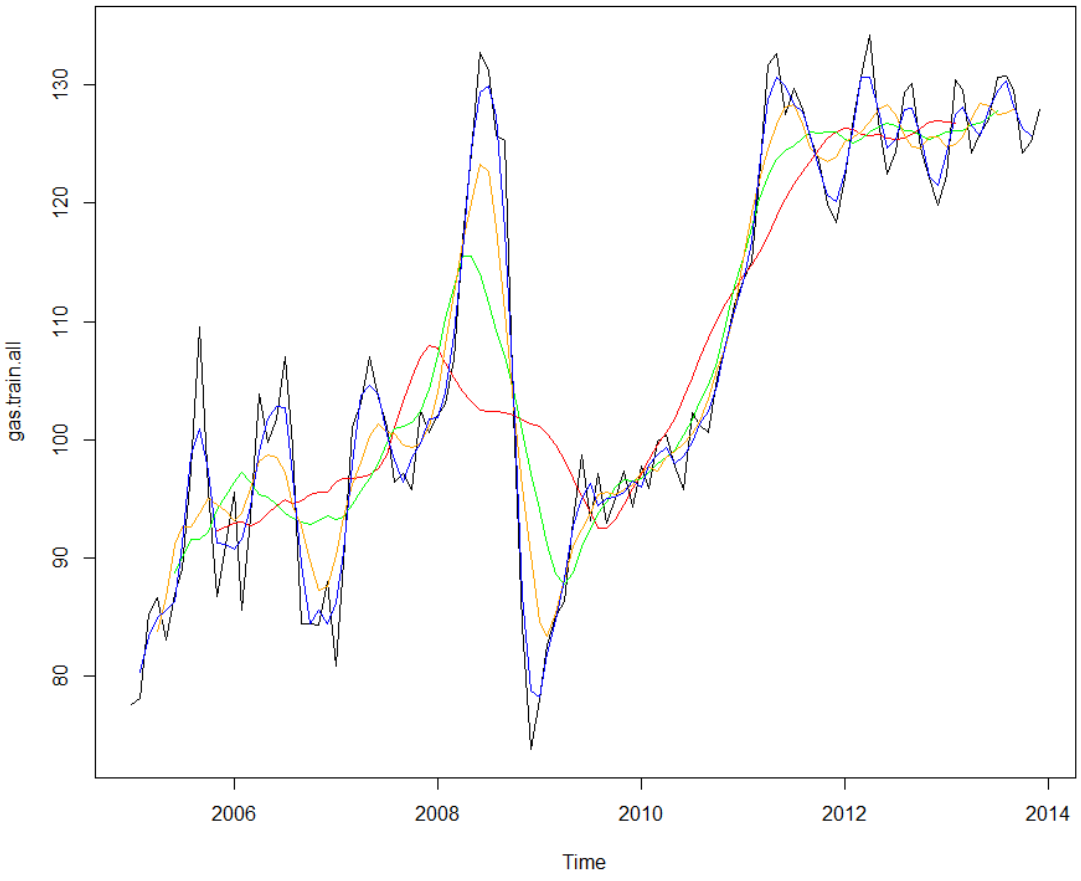


Figure 13: Various MA Filters for Training Set (Red -> q=10; Green -> q=5; Orange -. q=3; Blue -> q=1)

We next apply Simple Exponential Smoothing (SES) to our data. We predicted some seasonality in the data earlier, so we expect this model will not be too useful. It turns out that the fitted model is not very useful at all, as shown in Figure 14. The optimal α fitted by R was very close to one and we have the fitted values being the original values one step behind. Therefore, our model puts all of its emphasis on the data, making predictions futile.

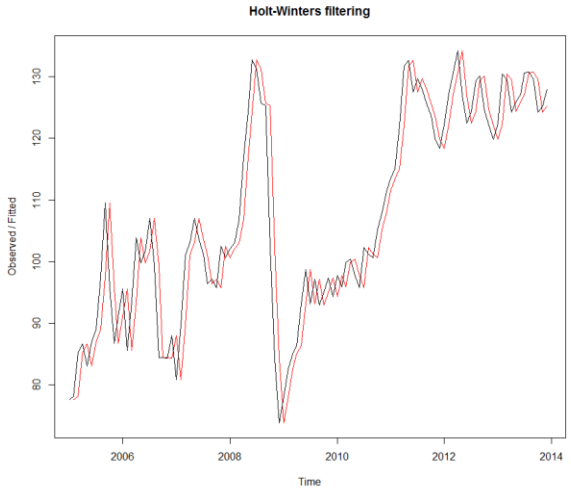


Figure 14: Fitted SES Model

In fact, Double Exponential Smoothing (DES) does not alleviate this problem. We get a very similar fitted model when we fit our data to such a model, as seen in Figure 15. In fact, R chose $\alpha = 1, \beta = 0$ as the optimal parameters, leaving us with the same problem as SES, in that the model relies completely on our data to predict values, so we cannot expect decent results if we wanted to forecast the future.

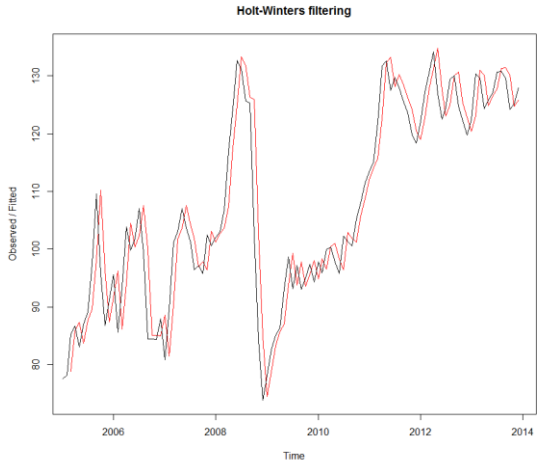


Figure 15: Fitted DES Model

As mentioned earlier, we stated that there are potential seasonality patterns in our data, which might lead to our SES and DES models failing. So, we would like to try both Additive and Multiplicative Holt-Winters models.

We first try the Additive Holt-Winters model, in which the fitted model is displayed in Figure 16. We can see that this model already does a better job at matching our data than our best regression model, without matching it exactly. We perform residual diagnostics, which can be seen in Figure 17. For this model, the ACF graph and the PACF graph tell us that there are minor correlations in the errors, which can be seen in the seemingly random small spikes. This might be coincidence, but it could indicate minor correlation. By looking at the QQ plot, the normality assumption is violated because of the heavy left tail. However, the residuals versus time plot indicates randomness among residuals, although the two big negative peaks are concerning.

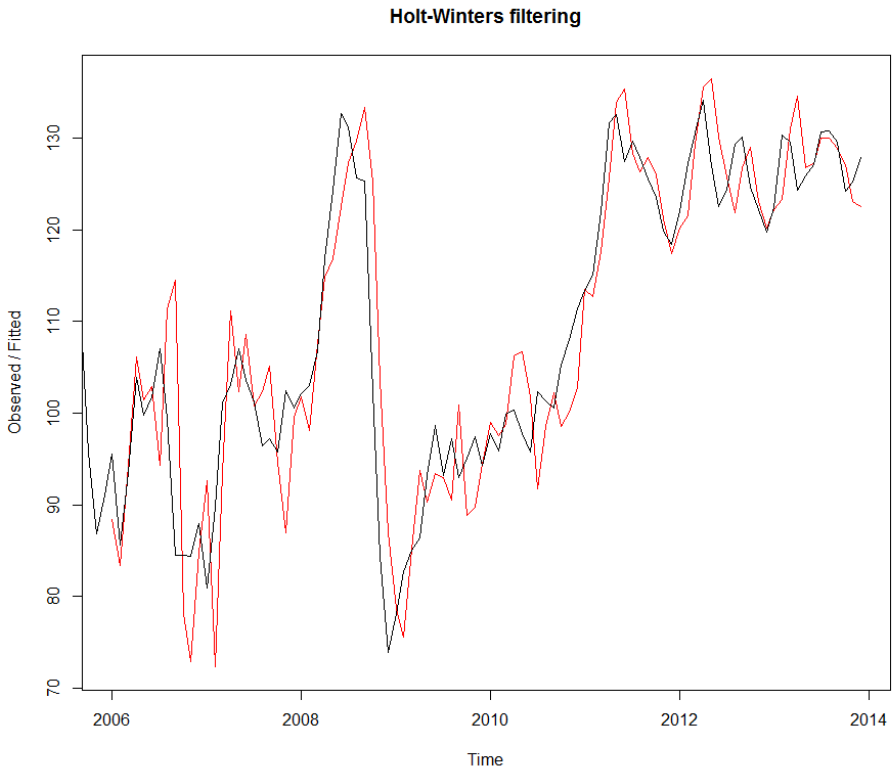


Figure 16: Fitted Additive Holt-Winters Model

To compare, we then fit the Multiplicative Holt-Winters model to our training set, as seen in Figure 18. This fit is nearly identical to our Additive model's fit. To see if there are any underlying problems with the residuals, we perform residual diagnostics, as seen in Figure 19. These residuals plots are nearly identical as well, so all the above comments apply.

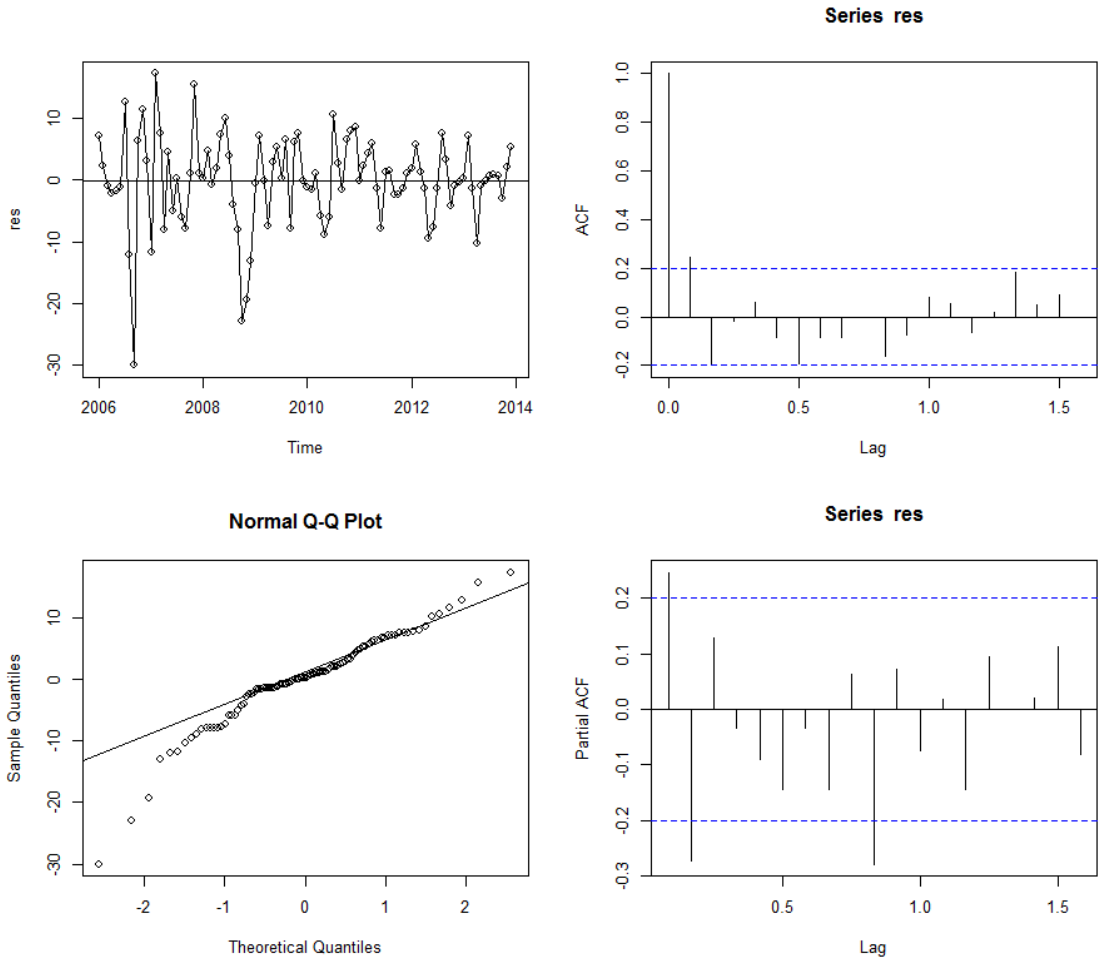


Figure 17: Residual Diagnostics for Additive Holt-Winters (Figure 16)

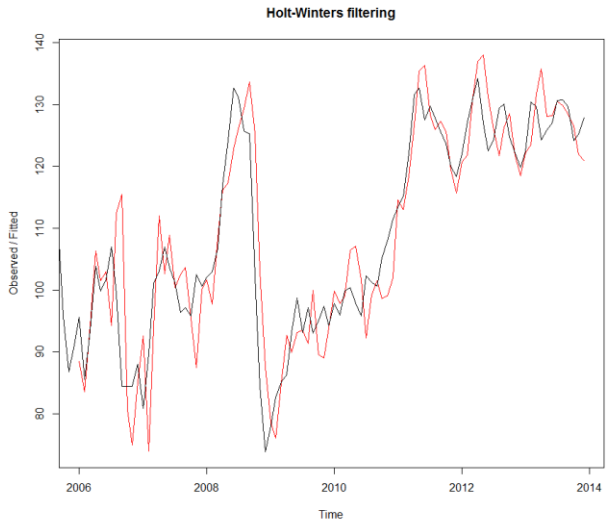


Figure 18: Fitted Multiplicative Holt-Winters Model

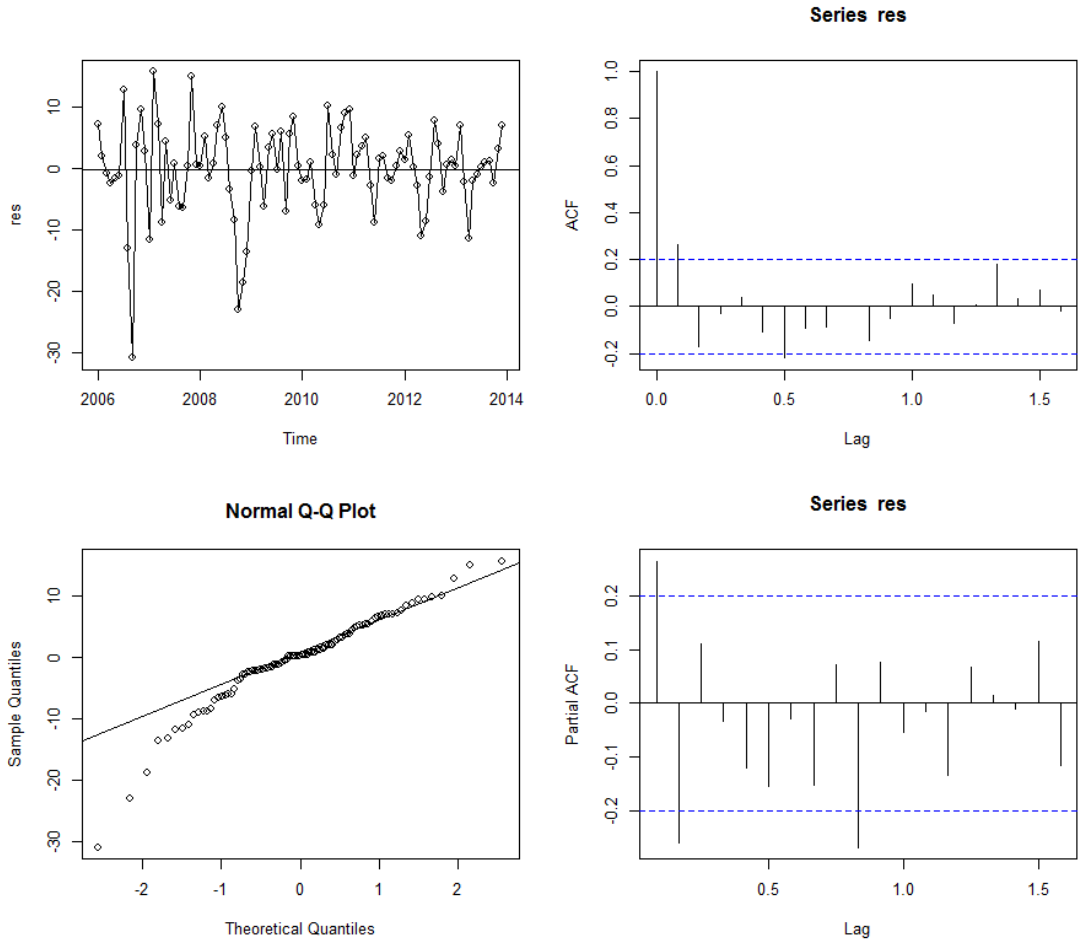


Figure 19: Residual Diagnostics for Multiplicative Holt-Winters (Figure 18)

It is clear that from our analysis that the Additive and Multiplicative Holt-Winters models are the best models we've found in our smoothing analysis, but now the question remains: which is better-suited for prediction? In order to test this, we will look at how each model predicts our testing set, and look at each model's PRESS value. Figure 20 shows how well the Additive model performed, and Figure 21 displays the Multiplicative model's performance, with the corresponding PRESS values in the captions. We see that the testing set falls under the prediction intervals of each model, but that since the Additive model has a lower PRESS value, it does a slightly better job of forecasting in this case.

Therefore, we choose our best smoothing method model to be our Additive Holt-Winter model. In the conclusion of this report, we will test the strength of it by forecasting gas prices for 2015 and compare its qualities to the other strong models we choose.

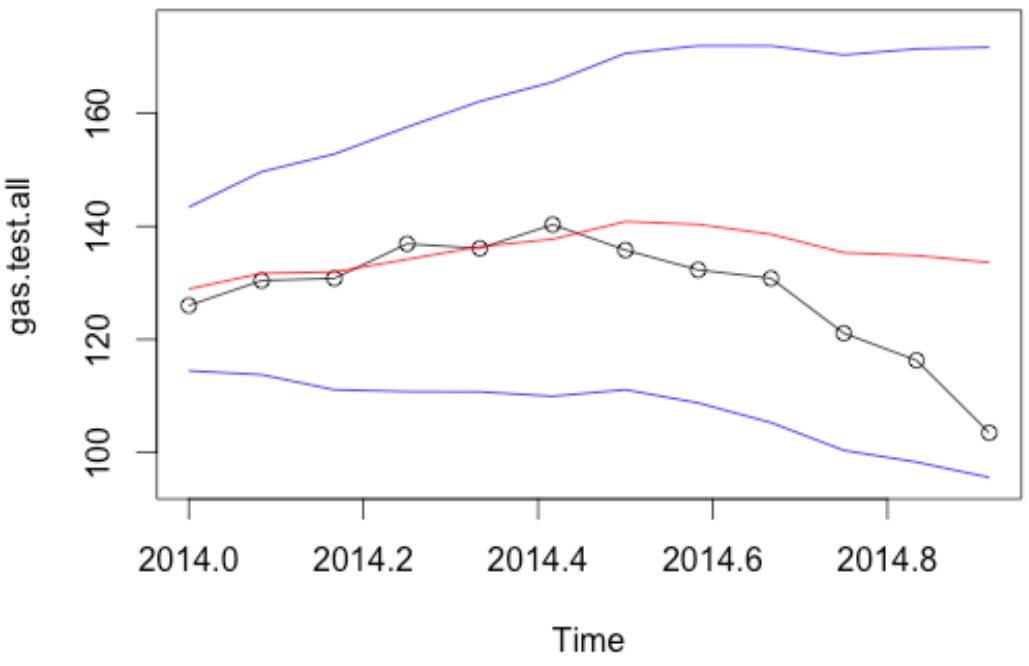


Figure 20: Additive Holt-Winters Model Predictions Compared to Testing Set (PRESS = 1631.236)

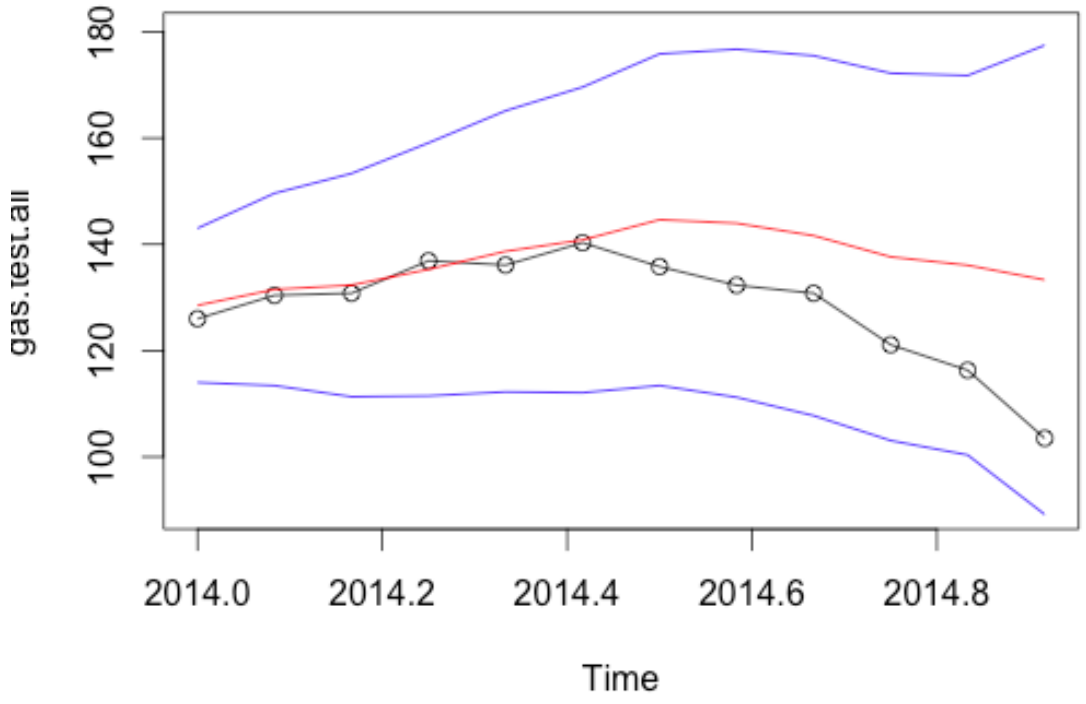


Figure 21: Multiplicative Holt-Winters Model Predictions Compared to Testing Set (PRESS = 1904.185)

Box-Jenkins Analysis

After considering the smoothing methods, we also like to consider applying other modelling methods, namely Box-Jenkins models.

In order to fit the models efficiently and correctly, we first analyze the data plot and the ACF graph. As shown from the data plot (Figure 1), and the nature of our data set, we can expect a monthly seasonal trend, so we can anticipate the SARIMA model to be a strong candidate. Along with that, from the ACF graph of our training set (Figure 22), shows significant correlation, which suggests differencing, will be needed to make our data stationary. Thus, the ARMA model is not considered, since our data is not stationary. Furthermore, from the ACF graph, we see no sign of a white noise pattern, so we can also eliminate the option of employing ARCH/GARCH models for our data.

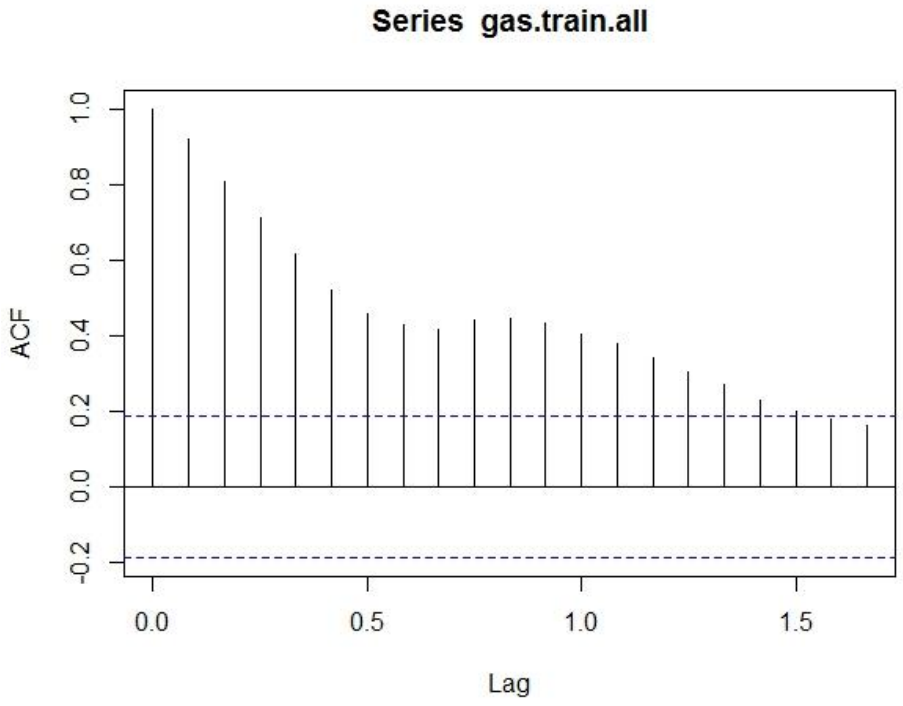


Figure 22: ACF of Data Training Set

We initially begin fitting an ARIMA model, and then we will try a SARIMA model to determine if there's any difference in the models, which indicates seasonality in our data set.

We start off with a single ordinary difference, which gives a data plot observed in Figure 23. The ACF (Figure 24) and PACF (Figure 25) graphs support the removal of trend, as the data now appears more stationary. Note that we can see a small seasonal trend within the significant bounds of these plots.

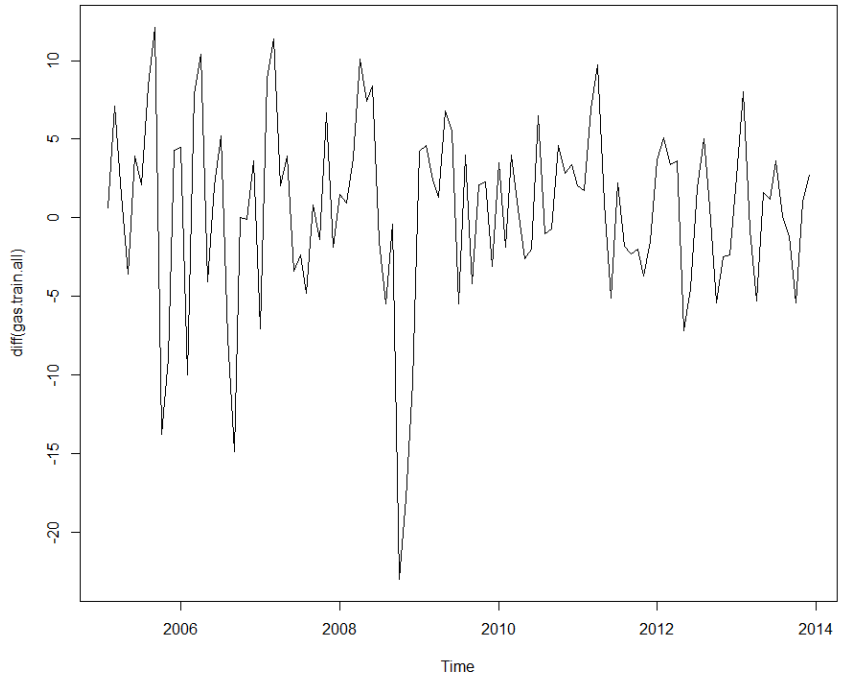


Figure 23: Time series after one ordinary difference

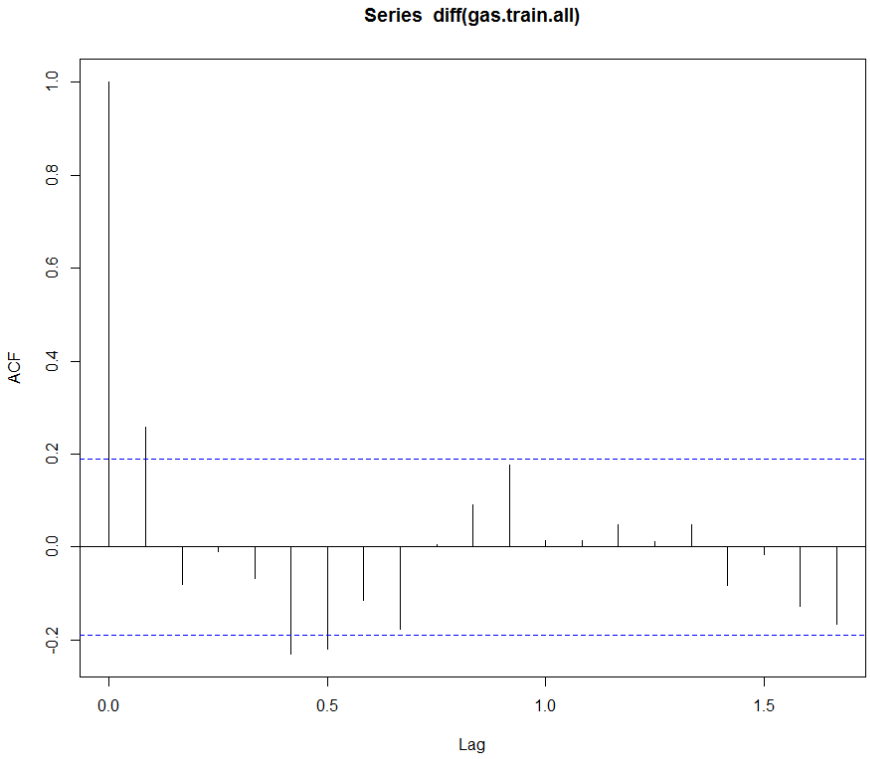


Figure 24: ACF of time series after one ordinary difference

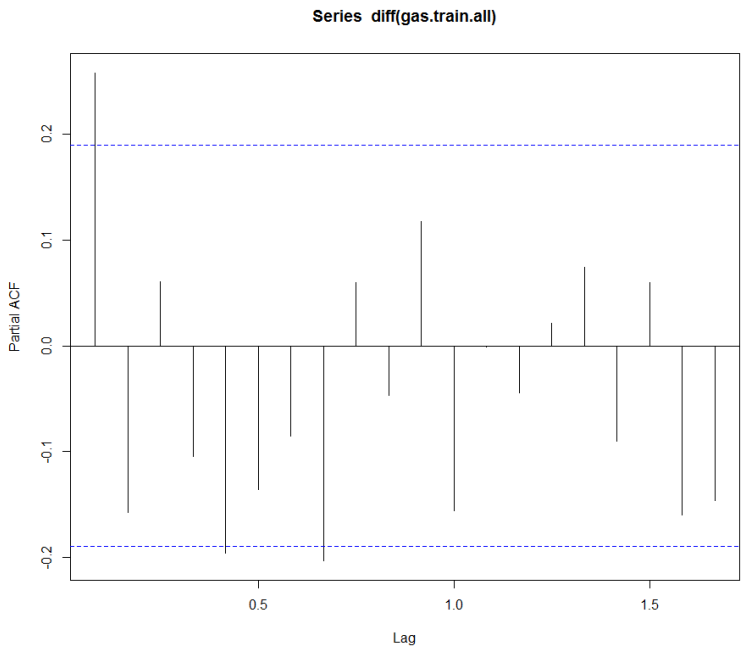


Figure 25: PACF for time series after one ordinary difference

Judging from the ACF and the PACF graph, we hypothesized that ARIMA(0,1,1) is probably the strongest candidate for the one initial spike in the graphs. We proceed to try combinations of 0 and 1 for parameters p and q. For completion, we also ordinarily differenced it twice, and did the same trial and error process. The result did not suggest that we needed to difference the data twice (see Appendix C-1). Finally, we compare the different models (Appendix C-2) with AIC values that lead to the conclusion that our hypothesis is indeed correct. The ARIMA(0,1,1) model is the best of the ARIMA models.

Hence, we move on to test the residuals of the model to ensure the viability (Figure 26). We can see that the residuals are randomly scattered with a few outliers. Furthermore, we observe the QQ plot being very close to linear, albeit with a heavy-tail on the left-end, which suggests that the residuals are potentially normal. Lastly, the ACF and PACF show minimal correlation among residuals, although there a couple spikes in both plots that pass the suggested bounds.

Next, we find a SARIMA model that fits our data. As mentioned earlier, we suspected SARIMA is a better fit for the data. Therefore, we begin with one seasonal difference. We plotted the differenced data plot and the ACF, PACF (Figure 27). The data plot shows a remaining seasonal trend while the ACF suggested the data remained non-stationary. We, therefore, have to difference our time series once more.

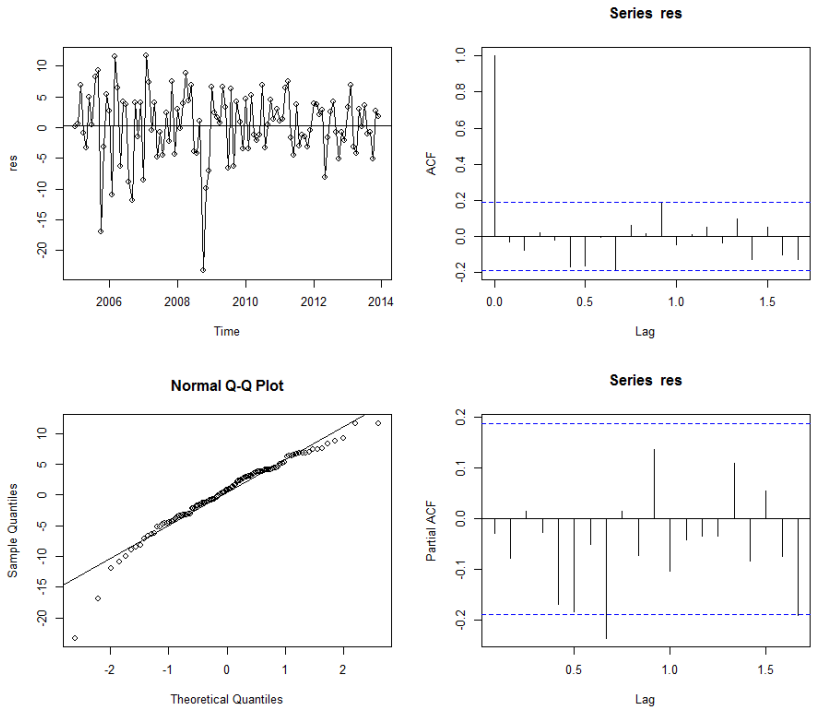


Figure 26: Residual Plots for ARIMA(0,1,1)

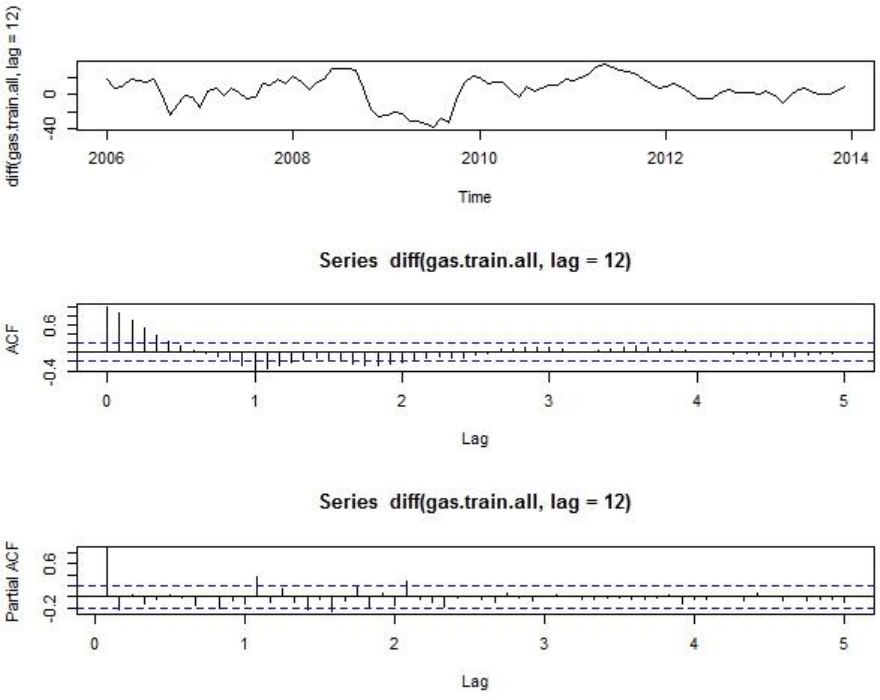


Figure 27: Seasonally differenced time series, its ACF and PACF

By differencing it once seasonally and ordinarily, we have reached stationary data (Figure 28). The data plot seems to have no trend, and ACF, PACF showed no correlation left behind. Hence, we may move on to the next step, estimating parameters P and Q.

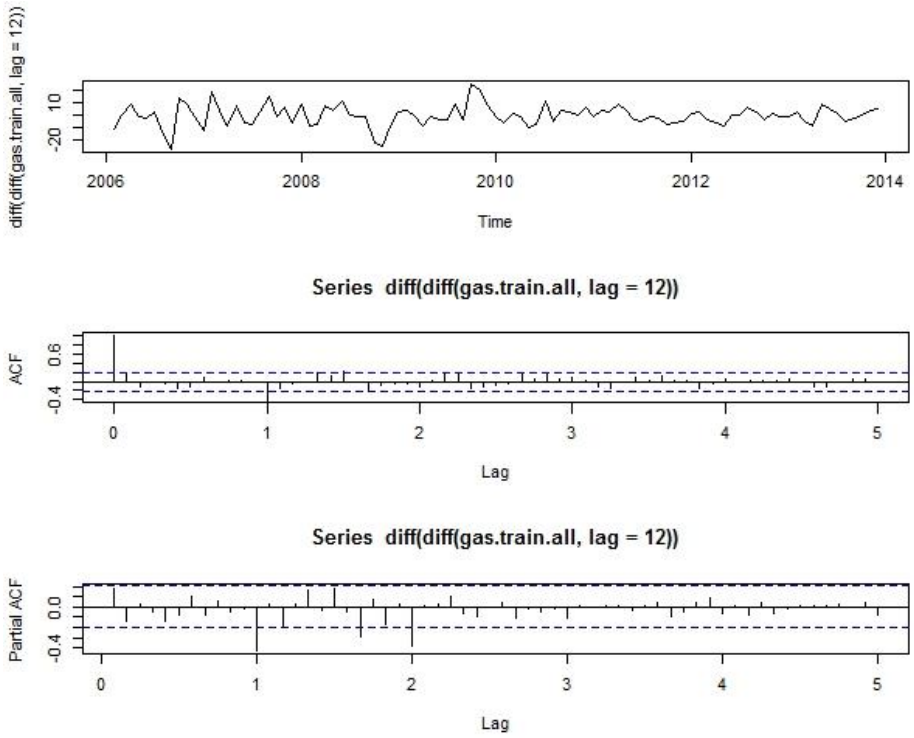


Figure 28: Differenced time series (d=1, D=1), and its ACF, PACF

From observing the ACF, we are sure that Q is 1 for the only spike at lag 12. Then, from trial and error (with p and q temporarily set to 0) and comparing the AIC values, we believe that P is 2 and Q is 1 (Appendix C-3). The next step would be to find out the other parameters, p and q. From ACF and PACF alone, it was very hard to figure out the values of p and q, so we estimate them by trial and error (Appendix C-4), knowing that they can't be too large for the amount and magnitude of spikes shown. We compared the AIC values (Appendix C-4) to find out the best model appeared to be SARIMA(0,1,2)x$(2,1,1)_{12}$.

Hence, we need to check the residuals through diagnostics (Figure 29). Similar to the ARIMA(0,1,1) model, the plots and graphs support that residuals are uncorrelated and normal, even though the QQ plot still has a heavy lower tail.

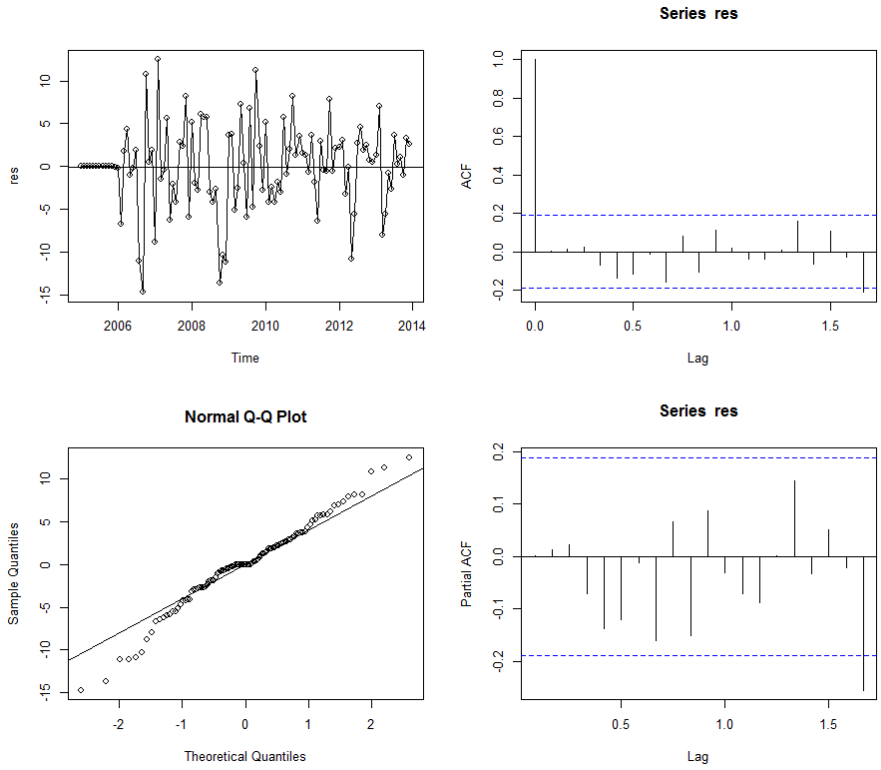


Figure 29: Residual Plots for SARIMA(0,1,2)x(2,1,1) $_{12}$

With the two final models from both ARIMA and SARIMA categories, we use AIC values for comparison to conclude the best model from the Box-Jenkins approach.

```
> gas.train.all.diff1ma1$aic
[1] 677.7132
> # SARIMA(0,1,2)x(2,1,1)_12
> gas.train.all.diff1sarma0221$aic
[1] 623.6878
```

By choosing the lower AIC value, we predict the testing set with the better model - SARIMA(0,1,2)x(2,1,1)$_{12}$. We construct a 95% prediction interval and predict the data for year 2014 from the SARIMA(0,1,2)x(2,1,1)$_{12}$ model (Figure 30). We can see although the predicted data (red) falls within the prediction interval (blue) entirely, the intervals become wider with time. To quantify, we calculated the PRESS statistic. As shown above, the PRESS, 2096.767, is rather large compared to the best models from other two approaches. Therefore, the Box-Jenkins approach is not the best approach.

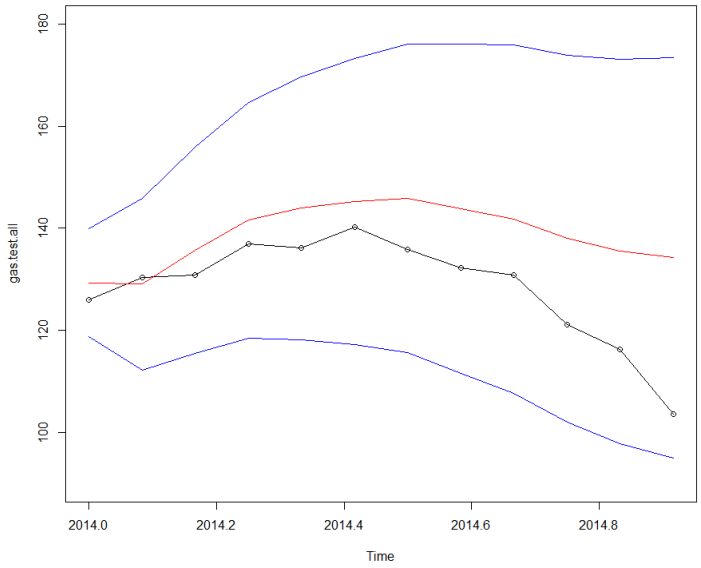


Figure 30: SARIMA(0,1,2)x(2,1,1)$_{12}$ Prediction Intervals for Testing Set, PRESS = 2096.767

Statistical Conclusions

Out of the three models, we narrow down to two potential candidates. The first candidate that we have is the regression model because it has the lowest PRESS value and strong long-term general trend. We also have Additive Holt Winters model because it is great for short term prediction. The SARIMA(0,1,2)x(2,1,1)$_{12}$ model does not compete with the Additive Holt-Winters model as they both having increasing prediction intervals, but the Additive model has a lower PRESS value, which is crucial since we are trying to predict future gas prices. However, it is worth noting that our SARIMA model has the best residuals of the three models. We summarize the qualities of our three best models in Table 2.

Table 2: Model Comparison Summary

	Regression	**Additive Holt Winters**	**SARIMA(0,1,2)x(2,1,1)$_{12}$**
PRESS	713.89	1631	2096
Fits	Reasonable	Excellent	Excellent
Normality of Residuals	Poor	Poor	Good
Random Residuals	Mediocre	Good	Excellent
ACF	Extreme Correlation (Not Stationary)	Good, but some correlation exists	Uncorrelated residuals
PACF	Good	Good, but some correlation exists	Uncorrelated residuals
Width of Prediction Interval	Narrow	Increases with time	Increases with time

To help decide between the regression and Additive Holt-Winters, we forecast gas price of 2015 with both models. We first observe the prediction intervals of 2015 for our regression model in Figure 31. Note that we re-fitted our best regression model (seasonal indicators with linear, quadratic and sinusoidal time component) to our entire data set and forecasted. We can see that our predictions fall completely out of range for what gas prices have actually been doing for the year (on a slow incline back up). Therefore, we shall look at our other model and hope that it provides better predictions that account for such variation.

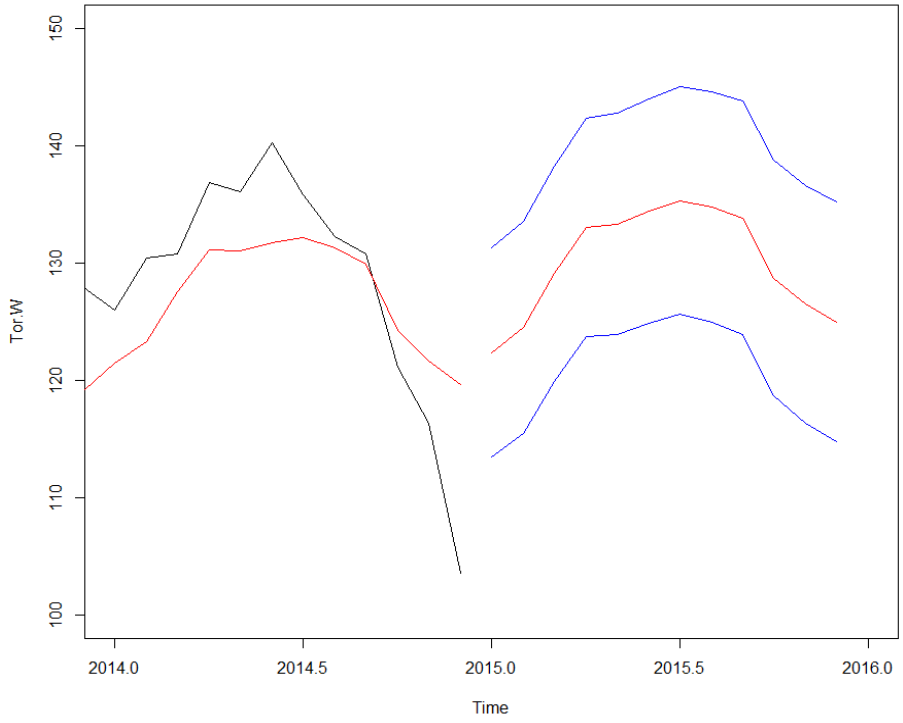


Figure 31: Forecasting 2015 gas prices using best regression model

We apply the same process to the Additive Holt-Winters model refitted to the entire data set, and thankfully have some better results, as seen in Figure 32. We can see that the actual prices do not have to change much to attain our predicted pump price for January 2015, and the intervals are wider to account for various sudden price changes.

Since we are more concerned with short term goals and objectives, we choose the Additive Holt-Winters model of these two final candidates. Although it is the model with the higher PRESS value, it is a much better fit to the actual data, and it accounts for variance in gas price a lot better.

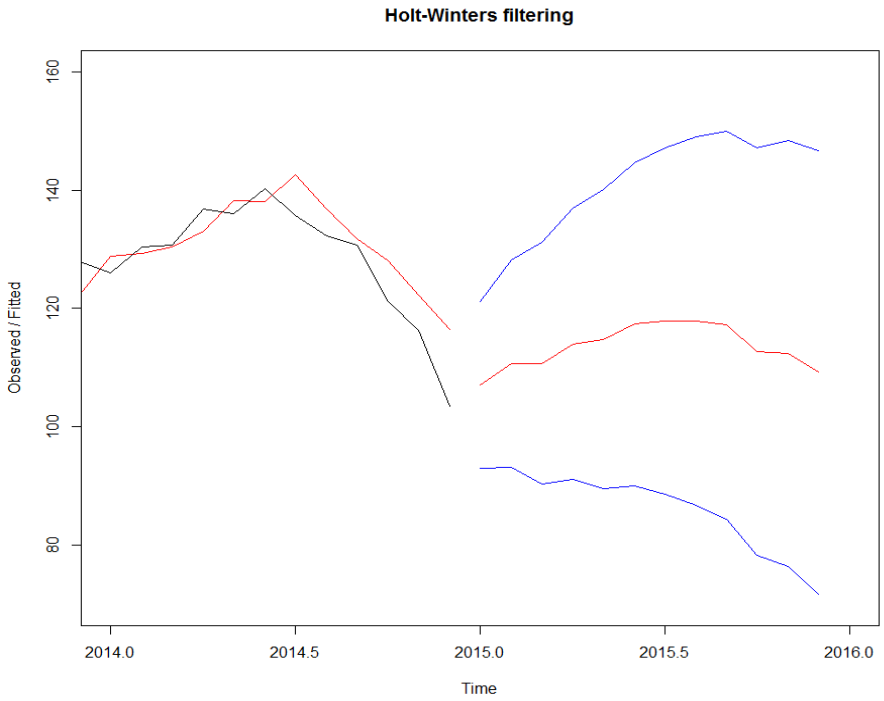


Figure 32: Forecasting 2015 gas prices using Additive Holt-Winters model

Conclusion to Problem

Based on the Additive Holt Winters model, the gas price for 2015 in West Toronto will peak around $1.18/L in the summer and will drop to $1.09/L by the year end. However, we expect variation with the peaks and drops of the gas price, so it will be necessary to update the model every month to get a more accurate prediction. We can use the model to predict gas prices in other parts of Ontario.

It is worth noting that even though this is the best model we found, we only considered time as an external factor. It may be useful in the future to perform a detailed analysis on potential factors (such as US exchange rate, local demand, driving population etc.) that impact gas prices locally in order to better model these sudden peaks and valleys in the data.

Appendix A – Data Set

Here is our original data set for reference. It was obtained from the Ontario Ministry of Energy (http://www.energy.gov.on.ca/en/fuel-prices/). We pulled the average price data from 2005 to 2014, representing price in cents. What follows is this data set in sorted order by time.

Jan-05	77.6	Jul-07	101.2	Jan-10	97.8	Jul-12	124.4
Feb-05	78.2	Aug-07	96.4	Feb-10	95.9	Aug-12	129.4
Mar-05	85.3	Sep-07	97.2	Mar-10	99.9	Sep-12	130.1
Apr-05	86.7	Oct-07	95.8	Apr-10	100.4	Oct-12	124.7
May-05	83.1	Nov-07	102.5	May-10	97.8	Nov-12	122.2
Jun-05	87	Dec-07	100.6	Jun-10	95.8	Dec-12	119.8
Jul-05	89.1	Jan-08	102.1	Jul-10	102.3	Jan-13	122.4
Aug-05	97.5	Feb-08	103	Aug-10	101.3	Feb-13	130.4
Sep-05	109.6	Mar-08	106.8	Sep-10	100.6	Mar-13	129.6
Oct-05	95.8	Apr-08	116.9	Oct-10	105.2	Apr-13	124.3
Nov-05	86.8	May-08	124.3	Nov-10	108	May-13	125.9
Dec-05	91.1	Jun-08	132.7	Dec-10	111.4	Jun-13	127.1
Jan-06	95.6	Jul-08	131.2	Jan-11	113.4	Jul-13	130.7
Feb-06	85.6	Aug-08	125.7	Feb-11	115.1	Aug-13	130.8
Mar-06	93.5	Sep-08	125.3	Mar-11	122	Sep-13	129.6
Apr-06	103.9	Oct-08	102.3	Apr-11	131.7	Oct-13	124.2
May-06	99.8	Nov-08	84.4	May-11	132.6	Nov-13	125.2
Jun-06	101.8	Dec-08	73.9	Jun-11	127.5	Dec-13	127.9
Jul-06	107	Jan-09	78.1	Jul-11	129.7	Jan-14	126
Aug-06	99.4	Feb-09	82.7	Aug-11	127.9	Feb-14	130.4
Sep-06	84.5	Mar-09	85.1	Sep-11	125.6	Mar-14	130.8
Oct-06	84.5	Apr-09	86.4	Oct-11	123.6	Apr-14	136.9
Nov-06	84.4	May-09	93.2	Nov-11	119.9	May-14	136.1
Dec-06	88	Jun-09	98.7	Dec-11	118.4	Jun-14	140.3
Jan-07	80.9	Jul-09	93.2	Jan-12	122.1	Jul-14	135.8
Feb-07	89.7	Aug-09	97.2	Feb-12	127.2	Aug-14	132.3
Mar-07	101.1	Sep-09	93	Mar-12	130.6	Sep-14	130.8
Apr-07	103.1	Oct-09	95.1	Apr-12	134.2	Oct-14	121.1
May-07	107	Nov-09	97.4	May-12	127	Nov-14	116.3
Jun-07	103.6	Dec-09	94.3	Jun-12	122.5	Dec-14	103.5

Appendix B – R Code and Output Combined

```
# Read in data file (sorted from original from oldest data to newest)
gas <- read.csv("~/Winter 2015/STAT 443/Project/gas.txt")
library("tseries")
# For the Runs Test, import lawstat library
library("lawstat")

# Diana's function for residual diagnostics
resdiags <- function(res) # you give this function a vector containing residuals from a model, as
well as vector of fitteds
{
  par(mfcol=c(2,2)) # splits the view to show 4 plots
  ts.plot(res) # time series plot of residuals
  points(res) # points to make counting runs easier
  abline(h=mean(res)) # mean line
  qqnorm(res) #qq plot
  qqline(res)
  acf(res) #acf
  acf(res, type="partial") #pacf
}

# Data broken up into different training and testing sets
# For all data, we take the last year as testing set
gas.all <- ts(gas, start = c(2005, 1), frequency = 12)
gas.train.all <- ts(gas[1:108,1], start = c(2005, 1), frequency = 12)
gas.test.all <- ts(gas[109:120,1], start = c(2014, 1), frequency = 12)

# Complete plot of all the data
plot(as.ts(gas.all))

# Plot of the training set data
plot(as.ts(gas.train.all))

# We investigate whether we should consider stablizing variance
# However, our log and square root plots don't really have an impact aside from scaling
par(mfcol=c(2,1))
plot(as.ts(gas.train.all))
plot(as.ts(log(gas.train.all)))
plot(as.ts(gas.train.all))
plot(as.ts((gas.train.all)^0.5))
par(mfcol=c(1,1))
# We conclude that we work with the data as is

################################# Regression Analysis

# Fitting pure seasonal model to training data
time <- time(gas.train.all) # Linear time term
months <- as.factor(cycle(time))
lm1 <- lm(gas.train.all~months)
points(time, lm1$fitted, col="red", type = "l")
resdiags(lm1$res)

# Fit a model with a linear time component and seasonality
lm2 <- lm(gas.train.all~time+months)
par(mfcol=c(1,1))
plot(as.ts(gas.train.all))
points(time, lm2$fitted, col="red", type = "l")
resdiags(lm2$res)

# Fit a model with seasonal indicators, linear time component, quadratic time component
time2 <- time^2 # For quadratic time term
lm3 <- lm(gas.train.all~time+time2+months)
par(mfcol=c(1,1))
plot(as.ts(gas.train.all))
points(time, lm3$fitted, col="red", type = "l")
resdiags(lm3$res)

# Fit a model with seasonal indicators, linear time component, quadratic time component, a
sinusoidal component
# The residual plot of lm3 looks like a sinusoidal pattern
```

```
lm4 <- lm(gas.train.all~time+time2+sin(time)+months)
par(mfcol=c(1,1))
plot(as.ts(gas.train.all))
points(time, lm4$fitted, col="red", type = "l")
resdiags(lm4$res)

# Fit a model with a linear time component, quadratic time component, a sinusoidal component
# The residual plot of lm3 looks like a sinusoidal pattern
lm5 <- lm(gas.train.all~time+time2+sin(time))
par(mfcol=c(1,1))
plot(as.ts(gas.train.all))
points(time, lm5$fitted, col="red", type = "l")
resdiags(lm5$res)
par(mfcol=c(1,1))

# We observe that in each case, the ACFs of the residuals indicate non-stationarity
# Therefore, we will not get any useful results by fitting an ARMA(p,q) model to the residuals of
our linear regression models
# To choose the "best" model, we compare some statistics
# First we compare adjusted R^2
summary(lm1)$adj.r.squared # Indicators
[1] -0.03985528
summary(lm2)$adj.r.squared # Indicators, Linear
[1] 0.5752806
summary(lm3)$adj.r.squared # Indicators, Linear, Quadratic
[1] 0.5916519
summary(lm4)$adj.r.squared # Indicators, Linear, Quadratic, Sin
[1] 0.6268124
summary(lm5)$adj.r.squared # Linear, Quadratic, Sin
[1] 0.6013368

# Next, standard error
summary(lm1)$sigma # Indicators
[1] 17.26759
summary(lm2)$sigma # Indicators, Linear
[1] 11.03561
summary(lm3)$sigma # Indicators, Linear, Quadratic
[1] 10.82083
summary(lm4)$sigma # Indicators, Linear, Quadratic, Sin
[1] 10.34449
summary(lm5)$sigma # Linear, Quadratic, Sin
[1] 10.69174

# Then, AIC
AIC(lm1) # Indicators
[1] 935.1177
AIC(lm2) # Indicators, Linear
[1] 839.2828
AIC(lm3) # Indicators, Linear, Quadratic
[1] 835.8946
AIC(lm4) # Indicators, Linear, Quadratic, Sin
[1] 827.0153
AIC(lm5) # Linear, Quadratic, Sin
[1] 824.2206

# Then, BIC
BIC(lm1) # Indicators
[1] 969.9855
BIC(lm2) # Indicators, Linear
[1] 876.8326
BIC(lm3) # Indicators, Linear, Quadratic
[1] 876.1265
BIC(lm4) # Indicators, Linear, Quadratic, Sin
[1] 869.9294
BIC(lm5) # Linear, Quadratic, Sin
[1] 837.6313

# lm 4 (Indicators, Linear, Quadratic, Sin) is either the best or second-best model for each
rating
# So, we will use this one for prediction
```

```
# We prepare the testing set for prediction intervals for seas4, our quadratic time, log,
seasonality
tim.test.all <- time(gas.test.all)
tim2.test.all <- tim.test.all^2
month.test <- as.factor(cycle(gas.test.all))
pred.test <- predict(lm4, newdata=list(time = tim.test.all, time2 = tim2.test.all, months =
month.test), se.fit=TRUE)
par(mfcol=c(1,1))
plot(gas.test.all, ylim=c(100, 145))
points(tim.test.all, gas.test.all)
points(tim.test.all, pred.test$fit, type="l", col="red")
points(tim.test.all, pred.test$fit+1.96*pred.test$se, type="l", col="blue")
points(tim.test.all, pred.test$fit-1.96*pred.test$se, type="l", col="blue")
# Observe that the last value in the testing set falls out of the prediction interval
# This is due to the unexpected fall of gas prices late 2014

# Our PRESS value
press.bestreg <-sum((gas.test.all - pred.test$fit)^2)
press.bestreg
[1] 713.8966

########################## Smoothing Analysis

# We first note that we don't attempt to model with a Slowly Drifting Mean Model
# This is because our model has too much variability in the mean, as seen from plot

##MA Filter: Although not useful for forecasting, we can use this to get an idea of the general
trend of the data
#General MA filter function
MAsmooth <- function(series, q)
{
  c = 1/(2*q+1) # constant to multiply by
  series.MA <- series # starting point Yt
  for(i in 1:q){ series.MA <- series.MA + lag(series, k=-i) + lag(series, k=i) } # adding Yt-i
and Yt+i to the averaged series for each i from 1 to q
  series.MA <- series.MA*c # multiplying by the constant
  return(series.MA)
}
#training set
plot(gas.train.all) # Data plot
lines(MAsmooth(gas.train.all,10),col="red") #q=10
lines(MAsmooth(gas.train.all,5),col="green") #q=5
lines(MAsmooth(gas.train.all,3),col="orange") #q=3
lines(MAsmooth(gas.train.all,1),col="blue") #q=1
#testing set
plot(gas.test.all) # Data plot
lines(MAsmooth(gas.test.all,3),col="orange") #q=3
lines(MAsmooth(gas.test.all,2),col="green") #q=2
lines(MAsmooth(gas.test.all,1),col="blue") #q=1

# SIMPLE EXPONENTIAL SMOOTHING
# We noted some seasonality in the data earlier, so we predict this won't be too useful
gas.ses <- HoltWinters(gas.train.all, beta = F, gamma = F)
gas.ses$alpha
[1] 0.9999339
plot(gas.ses)
# The alpha parameter is nearly 1, making this a useful model for prediction

# DOUBLE EXPONENTIAL SMOOTHING
# Here we encounter the same problem, as alpha = 1, beta = 0
gas.des <- HoltWinters(gas.train.all, gamma = F)
gas.des$alpha
alpha
  1
gas.des$beta
beta
   0
plot(gas.des)
# ADDITIVE SEASONAL HOLT WINTERS
# We anticipate this will be a better fit
gas.hwa <- HoltWinters(gas.train.all, seasonal = "add")
```

```
plot(gas.hwa)
resdiags(gas.train.all - gas.hwa$fitted[,1])
par(mfcol=c(1,1))

# MULTIPLICATIVE SEASONAL HOLT WINTERS
# We compare this to the additive model, very similar residuals
gas.hwm <- HoltWinters(gas.train.all, seasonal = "mult")
plot(gas.hwm)
resdiags(gas.train.all - gas.hwm$fitted[,1])
par(mfcol=c(1,1))

# We'll forecast with both Additive and Multiplicative models due to their apparent similarities
# First the additive
pred.hwa<-predict(gas.hwa, n.ahead=12, prediction.interval=TRUE)
plot(gas.test.all, ylim=c(95, 175))
points(tim.test.all, gas.test.all)
points(tim.test.all, pred.hwa[,1], type="l", col="red")
points(tim.test.all, pred.hwa[,2], type="l", col="blue")
points(tim.test.all, pred.hwa[,3], type="l", col="blue")
# Additive PRESS
press.hwa<-sum((gas.test.all - pred.hwa[,1])^2)
press.hwa
[1] 1631.236
# Then the multiplicative
pred.hwm<-predict(gas.hwm, n.ahead=12, prediction.interval=TRUE)
plot(gas.test.all, ylim=c(90, 180))
points(tim.test.all, gas.test.all)
points(tim.test.all, pred.hwm[,1], type="l", col="red")
points(tim.test.all, pred.hwm[,2], type="l", col="blue")
points(tim.test.all, pred.hwm[,3], type="l", col="blue")
# Multiplicative PRESS
press.hwm<-sum((gas.test.all - pred.hwm[,1])^2)
press.hwm
[1] 1904.185
# The Additive model has a lower PRESS value and more narrow prediction intervals
# Therefore this is deemed our best smoothing model

################### Box-Jenkins Models

## ARMA models
# Observe that we can't fit an ARMA model to our data
plot(gas.train.all) # clear upward trend and not stationary
acf(gas.train.all) # clearly correlated and not stationary

## ARIMA models

# We try differencing the data and see if we can fit an ARMA model to this differenced time
series
# We experiment with the number of differences and model types

#Diff=1
plot(diff(gas.train.all)) # trend seems to be removed
acf(diff(gas.train.all))
acf(diff(gas.train.all),type="partial")
#MA1, d=1
gas.train.all.diff1ma1<-arima(gas.train.all,order=c(0,1,1),method="ML")
gas.train.all.diff1ma1
Call:
arima(x = gas.train.all, order = c(0, 1, 1), method = "ML")

Coefficients:
         ma1
      0.3403
s.e.  0.0980

sigma^2 estimated as 31.73:  log likelihood = -336.86,  aic = 677.71
#AR1, d=1
gas.train.all.diff1ar1<-arima(gas.train.all,order=c(1,1,0),method="ML")
gas.train.all.diff1ar1
Call:
arima(x = gas.train.all, order = c(1, 1, 0), method = "ML")
```

```
Coefficients:
         ar1
      0.2600
s.e.  0.0929

sigma^2 estimated as 32.49:  log likelihood = -338.09,  aic = 680.19
#ARMA11, d=1
gas.train.all.diff1arma11<-arima(gas.train.all,order=c(1,1,1),method="ML")
gas.train.all.diff1arma11
Call:
arima(x = gas.train.all, order = c(1, 1, 1), method = "ML")

Coefficients:
          ar1     ma1
      -0.2197  0.5342
s.e.   0.2999  0.2643

sigma^2 estimated as 31.55:  log likelihood = -336.55,  aic = 679.1
#ARMA00 (White Noise), d=1
gas.train.all.diff1arma00 <-arima(gas.train.all, order=c(0,1,0), method="ML")
gas.train.all.diff1arma00
Call:
arima(x = gas.train.all, order = c(0, 1, 0), method = "ML")


sigma^2 estimated as 34.89:  log likelihood = -341.87,  aic = 685.73
##Diff=2
plot(diff(gas.train.all,difference=2))
acf(diff(gas.train.all,difference=2))
acf(diff(gas.train.all,difference=2),type="partial")
#MA1, d=2
gas.train.all.diff2ma1<-arima(gas.train.all,order=c(0,2,1),method="ML")
gas.train.all.diff2ma1
Call:
arima(x = gas.train.all, order = c(0, 2, 1), method = "ML")

Coefficients:
          ma1
      -1.0000
s.e.   0.0262

sigma^2 estimated as 34.99:  log likelihood = -341.17,  aic = 686.34
#AR1, d=2
gas.train.all.diff2ar1<-arima(gas.train.all,order=c(1,2,0),method="ML")
gas.train.all.diff2ar1
Call:
arima(x = gas.train.all, order = c(1, 2, 0), method = "ML")

Coefficients:
          ar1
      -0.2715
s.e.   0.0933

sigma^2 estimated as 48.05:  log likelihood = -355.67,  aic = 715.35
#ARMA11, d=2
gas.train.all.diff2arma11<-arima(gas.train.all,order=c(1,2,1),method="ML")
gas.train.all.diff2arma11
Call:
arima(x = gas.train.all, order = c(1, 2, 1), method = "ML")

Coefficients:
         ar1      ma1
      0.2672  -1.0000
s.e.  0.0941   0.0247

sigma^2 estimated as 32.67:  log likelihood = -337.26,  aic = 680.52
#ARMA21, d=2
gas.train.all.diff2arma21<-arima(gas.train.all,order=c(2,2,1),method="ML")
gas.train.all.diff2arma21
Call:
```

```
arima(x = gas.train.all, order = c(2, 2, 1), method = "ML")

Coefficients:
         ar1     ar2      ma1
      0.3041  -0.149  -1.0000
s.e.  0.0959   0.096   0.0252

sigma^2 estimated as 31.85:  log likelihood = -336.07,  aic = 680.15
#ARMA12, d=2
gas.train.all.diff2arma12<-arima(gas.train.all,order=c(1,2,2),method="ML")
gas.train.all.diff2arma12
Call:
arima(x = gas.train.all, order = c(1, 2, 2), method = "ML")

Coefficients:
          ar1      ma1      ma2
      -0.2044  -0.4764  -0.5236
s.e.   0.2961   0.2620   0.2612

sigma^2 estimated as 31.71:  log likelihood = -335.78,  aic = 679.56
#ARMA22, d=2
gas.train.all.diff2arma22<-arima(gas.train.all,order=c(2,2,2),method="ML")
gas.train.all.diff2arma22
Call:
arima(x = gas.train.all, order = c(2, 2, 2), method = "ML")

Coefficients:
          ar1      ar2      ma1      ma2
      -0.1000  -0.0443  -0.5826  -0.4174
s.e.   0.5036   0.1933   0.4997   0.4992

sigma^2 estimated as 31.68:  log likelihood = -335.76,  aic = 681.52
# We extract the AICs for comparison
#MA1, d=1
gas.train.all.diff1ma1$aic
[1] 677.7132
#AR1, d=1
gas.train.all.diff1ar1$aic
[1] 680.1855
#ARMA11, d=1
gas.train.all.diff1arma11$aic
[1] 679.0984
#ARMA00 (White Noise), d=1
gas.train.all.diff1arma00$aic
[1] 685.732
#MA1, d=2
gas.train.all.diff2ma1$aic
[1] 686.3365
#AR1, d=2
gas.train.all.diff2ar1$aic
[1] 715.3498
#ARMA11, d=2
gas.train.all.diff2arma11$aic
[1] 680.5219
#ARMA21, d=2
gas.train.all.diff2arma21$aic
[1] 680.1488
#ARMA12, d=2
gas.train.all.diff2arma12$aic
[1] 679.5561
#ARMA22, d=2
gas.train.all.diff2arma22$aic
[1] 681.5166
#According to AIC, we should use at MA1, d=1
resdiags(gas.train.all.diff1ma1$res)
par(mfcol=c(1,1))
# So our best ARIMA model is ARIMA(0,1,1)

##SARIMA

# Our series isn't stationary after one seasonal difference, D = 1
```

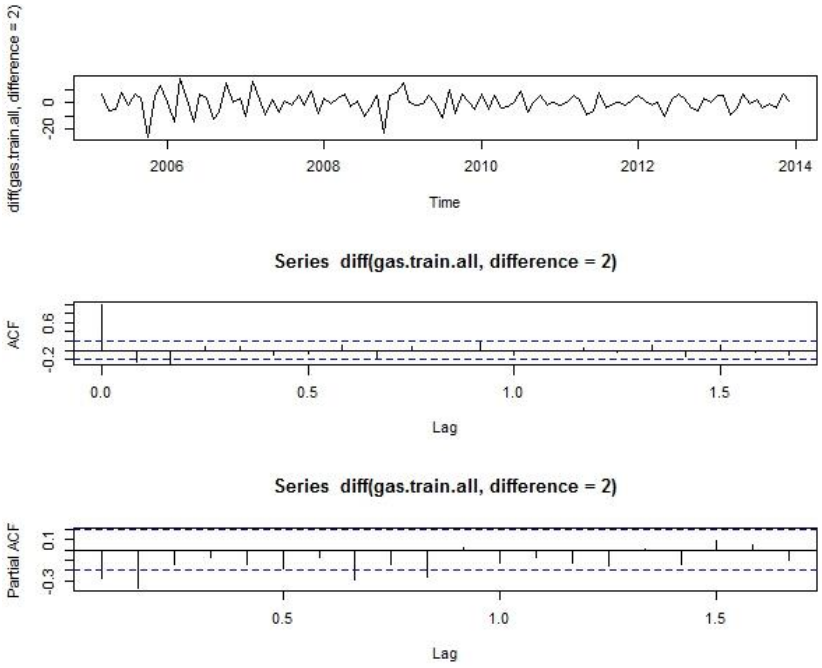
```
plot(diff(gas.train.all,lag=12))
acf(diff(gas.train.all,lag=12),lag.max=60)
acf(diff(gas.train.all,lag=12),lag.max=60,type="partial")
# From the dataplot, acf and pacf, we see that the series is not stationary hence we move on to
the following.
##diff 1 normal 1 seasonal

plot(diff(diff(gas.train.all,lag=12))) # plot normal diff1 sesonaldiff 1
acf(diff(diff(gas.train.all,lag=12)),lag.max=60)# acf
acf(diff(diff(gas.train.all,lag=12)),lag.max=60,type="partial") #pacf
# Acf and pacf shows stationarity.Hence, we can move on to fit the data into SARIMA models.

# Note: SARIMApqPQ for d, D means a fit to SARIMA(p,d,q)x(P,D,Q)

#SARIMA0021, d=1 D=1

gas.train.all.diff1sarma0021<-arima(gas.train.all,order=c(0,1,0),seasonal=list(order=c(2,1,1),
period=12),method="ML")
gas.train.all.diff1sarma0021$aic
[1] 627.507
#SARIMA0011, d=1 D=1
gas.train.all.diff1sarma0011<-arima(gas.train.all,order=c(0,1,0),seasonal=list(order=c(1,1,1),
period=12),method="ML")
gas.train.all.diff1sarma0011$aic
[1] 629.1414
#SARIMA P=2 Q=1 is preferred with a lower AIC value
#Now we will determine p and q by trial and error.
#p=1 q=0
gas.train.all.diff1sarma1021<-arima(gas.train.all,order=c(1,1,0),seasonal=list(order=c(2,1,1),
period=12),method="ML")
gas.train.all.diff1sarma1021$aic
[1] 626.296
#p=0 q=1
gas.train.all.diff1sarma0121<-arima(gas.train.all,order=c(0,1,1),seasonal=list(order=c(2,1,1),
period=12),method="ML")
gas.train.all.diff1sarma0121$aic
[1] 624.5287
#p=2, q=0
gas.train.all.diff1sarma2021<-arima(gas.train.all,order=c(2,1,0),seasonal=list(order=c(2,1,1),
period=12),method="ML")
gas.train.all.diff1sarma2021$aic
[1] 625.3588
#p=0, q=2
gas.train.all.diff1sarma0221<-arima(gas.train.all,order=c(0,1,2),seasonal=list(order=c(2,1,1),
period=12),method="ML")
gas.train.all.diff1sarma0221$aic
[1] 623.6878
#p=0, q=3
gas.train.all.diff1sarma0321<-arima(gas.train.all,order=c(0,1,3),seasonal=list(order=c(2,1,1),
period=12),method="ML")
gas.train.all.diff1sarma0321$aic
[1] 625.5324
# By observation and trial/error, we conclude that SARIMA(0,1,2)x(2,1,1)_12 is the best model out
of all SARIMA models we try
# We analyze residuals
resdiags(gas.train.all.diff1sarma0221$res)
par(mfcol=c(1,1))
## ARCH/GARCH

acf(gas.train.all) # This plot does not resemble white noise, so ARCH/GARCH is not suitable alone
## So compare the AICs of our chosen ARIMA and SARIMA model
# ARIMA(0,1,1)
gas.train.all.diff1ma1$aic
[1] 677.7132
# SARIMA(0,1,2)x(2,1,1)_12
gas.train.all.diff1sarma0221$aic
[1] 623.6878
# We choose the SARIMA model. Now we forecast with it and compare its results to the testing set
plot(gas.test.all, ylim=c(90, 180))
points(tim.test.all, gas.test.all)
pred.sarima <- predict(gas.train.all.diff1sarma0221, n.ahead=12, se.fit=TRUE)
```

```
points(tim.test.all, pred.sarima$pred, type="l", col="red")
points(tim.test.all, pred.sarima$pred + 1.96*pred.sarima$se, type="l", col="blue")
points(tim.test.all, pred.sarima$pred - 1.96*pred.sarima$se, type="l", col="blue")
# Calculate its press
press.sarima <- sum((gas.test.all - pred.sarima$pred)^2)
press.sarima
[1] 2096.767

### CONCLUSION and FORECAST FUTURE
# We decide on our Additive Holt-Winters to our best model
# We forecast 2015
# First, we refit our model using the total set
gas.hwa.full <- HoltWinters(gas.all, seasonal = "add")
par(mfcol=c(1,1))
plot(gas.hwa.full, xlim=c(2014,2016), ylim=c(70,160))
pred.hwa2<-predict(gas.hwa.full, n.ahead=12, prediction.interval=TRUE)
points(tim.test.all+1, pred.hwa2[,1], type="l", col="red")
points(tim.test.all+1, pred.hwa2[,2], type="l", col="blue")
points(tim.test.all+1, pred.hwa2[,3], type="l", col="blue")

# We forecast 2015 with our regression model, refitted to all the data
time.all <- time(gas.all) # Linear time term
months.all <- as.factor(cycle(time.all))
time.all2 <- time.all^2 # For quadratic time term
lm4.full <- lm(gas.all~time.all+time.all2+sin(time.all)+months.all)
plot(gas.all, xlim=c(2014,2016), ylim=c(100,150))
points(time.all, lm4.full$fitted, col="red", type = "l")
# predictions
tim.future <- time(gas.test.all)+1
tim2.future <- tim.future^2
pred.future <- predict(lm4.full, newdata=list(time.all = tim.future, time.all2 = tim2.future,
months.all = month.test), se.fit=TRUE)
points(tim.future, pred.future$fit, type="l", col="red")
points(tim.future, pred.future$fit+1.96*pred.future$se, type="l", col="blue")
points(tim.future, pred.future$fit-1.96*pred.future$se, type="l", col="blue")

# Overall conclusion is that the Additive Holt Winters model is best
# But it should be updated as often as possible to get most accurate predictions
# As time passes, prediction intervals widen
```

The following C appendices are supplemental plots, and excerpts from the R Code to make it easier for the reader to identify the team's decision making process.

Appendix C-1

Ordinary differencing twice. The data plot and ACF, PACF graphs



Appendix C-2

Trial and error for ARIMA models

```
> #MA1, d=2
> gas.train.all.diff2ma1<-arima(gas.train.all,order=c(0,2,1),method="ML")
> #AR1, d=2
> gas.train.all.diff2ar1<-arima(gas.train.all,order=c(1,2,0),method="ML")
> #ARMA11, d=2
> gas.train.all.diff2arma11<-arima(gas.train.all,order=c(1,2,1),method="ML")
> #ARMA21, d=2
> gas.train.all.diff2arma21<-arima(gas.train.all,order=c(2,2,1),method="ML")
> #ARMA12, d=2
> gas.train.all.diff2arma12<-arima(gas.train.all,order=c(1,2,2),method="ML")
> #ARMA22, d=2
> gas.train.all.diff2arma22<-arima(gas.train.all,order=c(2,2,2),method="ML")
```

Comparing AIC values for ARIMA models

```
> # We extract the AICs for comparison
> #MA1, d=1
> gas.train.all.diff1ma1$aic
[1] 677.7132
> #AR1, d=1
> gas.train.all.diff1ar1$aic
[1] 680.1855
> #ARMA11, d=1
> gas.train.all.diff1arma11$aic
[1] 679.0984
> #ARMA00 (White Noise), d=1
> gas.train.all.diff1arma00$aic
[1] 685.732
> #MA1, d=2
> gas.train.all.diff2ma1$aic
[1] 686.3365
> #AR1, d=2
> gas.train.all.diff2ar1$aic
[1] 715.3498
> #ARMA11, d=2
> gas.train.all.diff2arma11$aic
[1] 680.5219
> #ARMA21, d=2
> gas.train.all.diff2arma21$aic
[1] 680.1488
> #ARMA12, d=2
> gas.train.all.diff2arma12$aic
[1] 679.5561
> #ARMA22, d=2
> gas.train.all.diff2arma22$aic
[1] 681.5166
```

Appendix C-3

Discovering P and Q through trial and error

```
> # Note: SARIMApqPQ for d, D means a fit to SARIMA(p,d,q)x(P,D,Q)
> #SARIMA0021, d=1 D=1
> gas.train.all.diff1sarma0021<-arima(gas.train.all,order=c(0,1,0),seasonal=list(order
=c(2,1,1), period=12),method="ML")
> gas.train.all.diff1sarma0021$aic
[1] 627.507
> #SARIMA0011, d=1 D=1
> gas.train.all.diff1sarma0011<-arima(gas.train.all,order=c(0,1,0),seasonal=list(order
=c(1,1,1), period=12),method="ML")
> gas.train.all.diff1sarma0011$aic
[1] 629.1414
> #SARIMA P=2 Q=1 is preferred with a lower AIC value
```

Appendix C-4

Discovering p and q through trial and error and comparing AIC values for SARIMA models

```
> #Now we will determine p and q by trial and error.
> #p=1 q=0
> gas.train.all.diff1sarma1021<-arima(gas.train.all,order=c(1,1,0),seasonal=list(order
=c(2,1,1), period=12),method="ML")
> gas.train.all.diff1sarma1021$aic
[1] 626.296
> #p=0 q=1
> gas.train.all.diff1sarma0121<-arima(gas.train.all,order=c(0,1,1),seasonal=list(order
=c(2,1,1), period=12),method="ML")
> gas.train.all.diff1sarma0121$aic
[1] 624.5287
> #p=2, q=0
> gas.train.all.diff1sarma2021<-arima(gas.train.all,order=c(2,1,0),seasonal=list(order
=c(2,1,1), period=12),method="ML")
> gas.train.all.diff1sarma2021$aic
[1] 625.3588
> #p=0, q=2
> gas.train.all.diff1sarma0221<-arima(gas.train.all,order=c(0,1,2),seasonal=list(order
=c(2,1,1), period=12),method="ML")
```

```
> gas.train.all.diff1sarma0221$aic
[1] 623.6878
> #p=0, q=3
> gas.train.all.diff1sarma0321<-arima(gas.train.all,order=c(0,1,3),seasonal=list(order
=c(2,1,1), period=12),method="ML")
> gas.train.all.diff1sarma0321$aic
[1] 625.5324
> # By observation and trial/error, we conclude that SARIMA(0,1,2)x(2,1,1)_12 is the b
est model out of all SARIMA models we try
```