Ontario Flu Trends

GROUP 2

Alaric Sin Hong Chow

Jennifer Luu

Lam Diep Nguyen

Nitharshan Nithianantharajah

Suchitra Ramlakhan

Stat 443 Final Project

April 3, 2017

Member Contributions

Alaric Sin Hong Chow • 20517917

Regression Analysis

Jennifer Luu • 20413026

Introduction

Presentation of Data

Lam Diep Nguyen • 20464564

Box-Jenkins Analysis

Nitharshan Nithianantharajah • 20383745

Smoothing Methods

Suchitra Ramlakhan • 20455631

Conclusion

Composition of Project

Table of Contents

List of Figures and Tables

Figures

Tables

1.0 Introduction

Influenza, also known as the flu, is a contagious disease that affects the nose, throat and lungs. Millions of Canadians become infected by the flu each year. Although most people recover from the flu, the severity of the illness can range from mild to extreme, including serious health complications or even death. The goal of this report is to analyze the historical flu trends to forecast future trends in order to bring public awareness to the rate of infection of influenza in hopes of encouraging both the federal and provincial governments to supply sufficient resources to combat and decrease the number of infections in Ontario.

Using approximately twelve years worth of historical data from 2003 to 2015, we will apply various statistical methods to the data to find a statistical relationship in the number of people who are infected by the flu in Ontario. We hypothesize the following:

- *Null hypothesis*: H_0 = There is a decrease in the number of people catching the flu
- *Alternative Hypothesis*: H_1 = The is an increase or no change in the number of people catching the flu

With our hypothesis in mind, we will apply three types of statistical methods to the data:

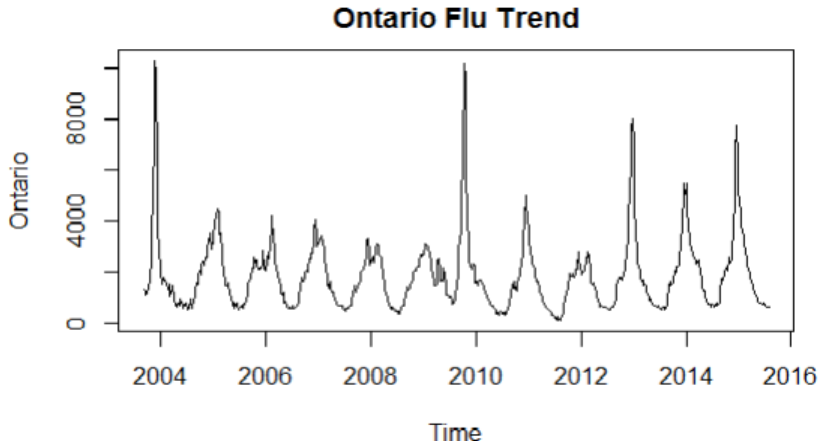
- Smoothing methods
- Regression analysis
- Box-Jenkins models

For each of the three methods, we will attempt to fit or model the data and check for the appropriateness of the fit. After checking and comparing the models from each of the methods, the best candidate model will be selected to forecast the values 155 weeks into the future.

2.0 Presentation of Data

The data that will be used in the analysis was collected and obtained by Google[1], and it contains the number of people who were infected by the flu on a weekly basis, starting from the week of September 28, 2003 to the week of August 9, 2015, presenting a total of 620 weeks of historical data to analyze The observed graph below provides a plot of the data (Figure 1).

Figure 2.1: Historical Ontario Flu Trend from 2003 to 2015



To perform a more efficient analysis, the data was decomposed into its trend, seasonal, and random components (Figure 2). From the seasonal component, it is evident that there are seasonal effects on the data – more specifically, on average, there are a higher number of people with the flu during the winter time than any other time of the year (Table 1). This seasonal effect appears to be stable each year. The trend component shows some consistency from 2004 to 2009, and there appears to be an increasing trend from 2012 to 2015. However, further analysis is required to provide a conclusive result on the flu trend. The random component shows the fluctuations in the data. A few outliers appeared in the data. Particularly, a significant outlier occurred at the end of 2009 that caused instability in the data; this instability was caused by the 2009 H1N1 flu pandemic in Canada.

Figure 2.2: Decomposition of the Ontario Flu Data

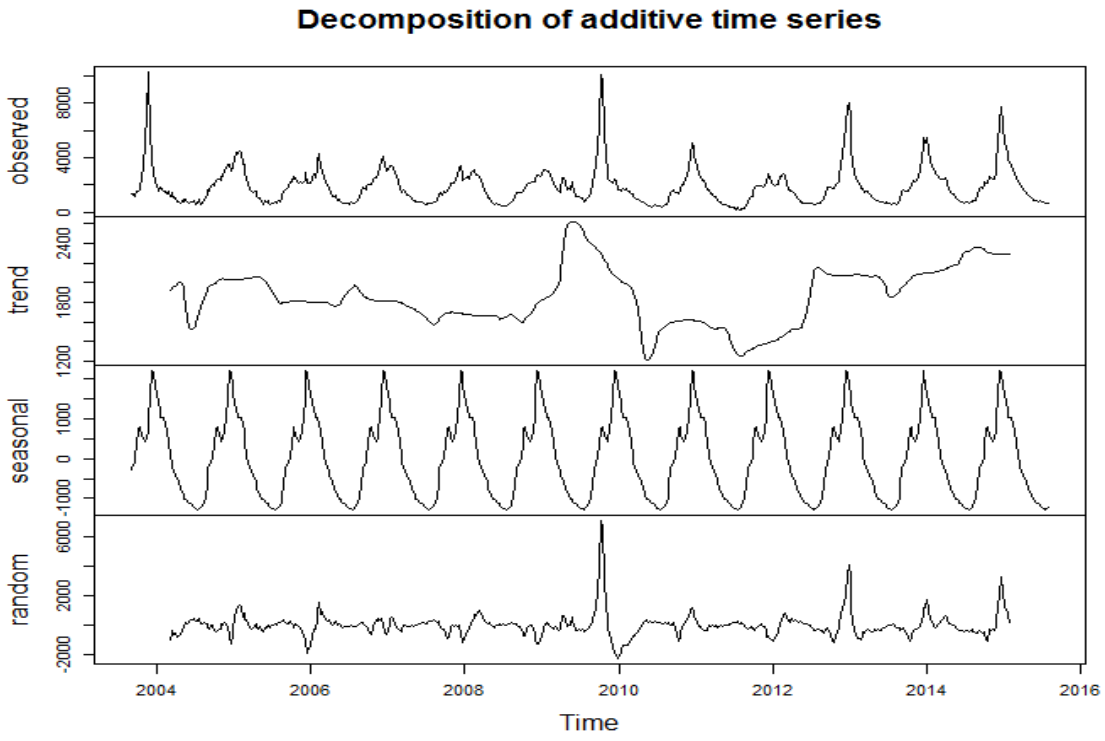


Table 2.1: Breakdown of Ontario Flu by Seasons

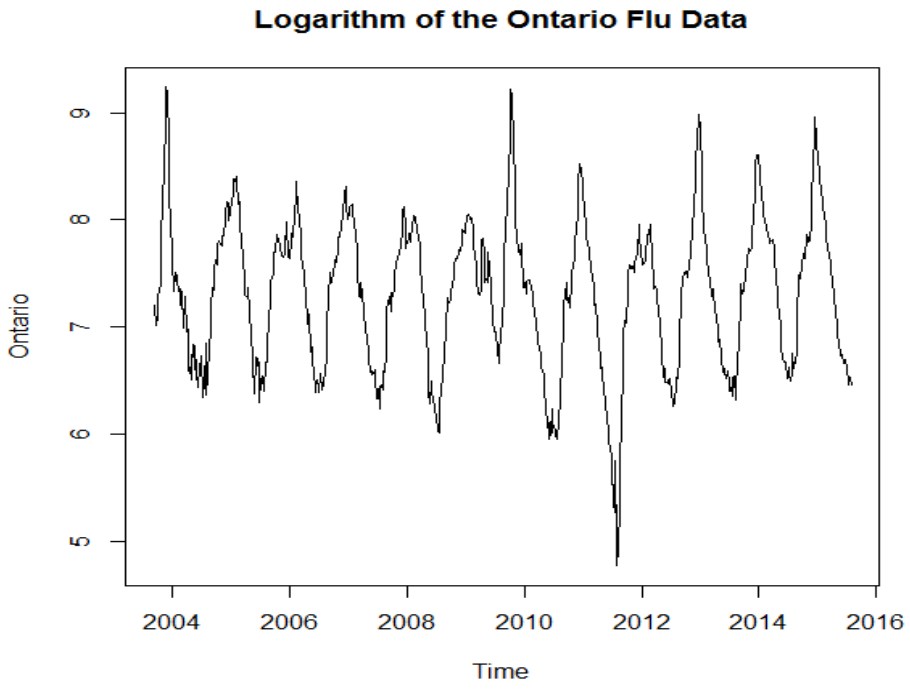
Sum of Ontario Years[1]	Seasons[2] Autumn	Winter	Spring	Summer	Total
2004	29,875	25,705	12,900	10,962	79,442
2005	28,109	46,629	17,855	10,690	103,283
2006	29,224	42,414	18,542	9,830	100,010
2007	23,827	39,195	14,540	9,291	86,853
2008	23,389	34,878	18,455	7,951	84,673
2009	60,237	32,992	24,538	18,137	135,904
2010	27,224	23,243	10,697	7,751	68,915
2011	22,417	36,256	10,777	4,055	73,505
2012	32,437	41,914	14,702	9,431	98,484
2013	28,512	48,622	13,862	9,589	100,585
2014	33,406	53,327	21,256	11,313	119,302
Total	**338,657**	**425,175**	**178,124**	**109,000**	**1,050,956**
% of Total	**32%**	**40%**	**17%**	**10%**	**100%**
Average[3]	**27,842**	**39,218**	**15,359**	**9,086**	**95,541**

Notes:

1. Years 2003 and 2015 were not included due to incomplete yearly data
2. For this report, each season is defined as:
   1. Autumn: Sep 21-Dec 20
   2. Winter: Dec 21-Mar 20
   3. Spring: Mar 21-Jun 20
   4. Summer: Jun 21-Sep
3. The average excludes 2009 due to the extreme abnormality in that year

A logarithm transformation was applied to the data to reduce the extreme volatility and improve the interpretability of the graph. From the post-transformation results (Figure 3), the plot shows less volatility and a consistent seasonal and trend in the data.

Figure 2.3: Logarithm of the Data



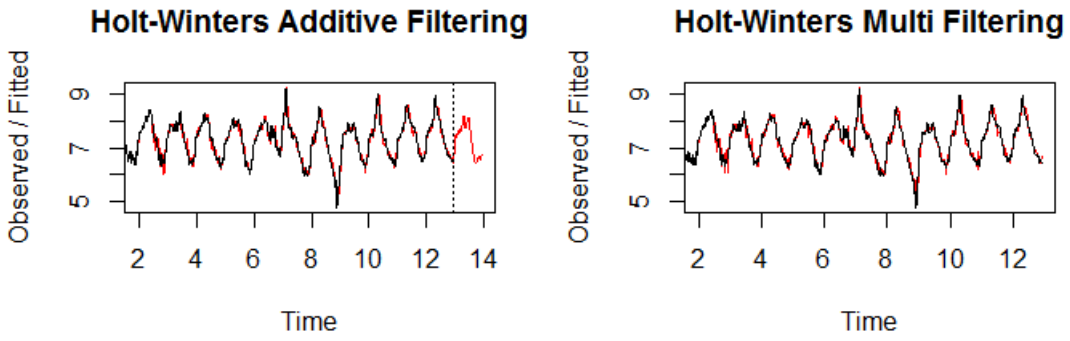
3.0 Data Analysis

In an attempt to test our hypothesis of decreasing flu patterns, the data was analyzed in 3 major ways. Firstly, the Holt-Winters smoothing methods (multiplicative and additive) were used to model the data and produce predictions. Secondly, various regression models were fitted to the data and the best ones were chosen to make forecasts. Lastly, a Box-Jenkins approach was employed in the same manner as the Regression Analysis. Upon completion of these 3 methodologies, a conclusion will be made on the best model, overall.

3.1 Smoothing Methods: Holt-Winters

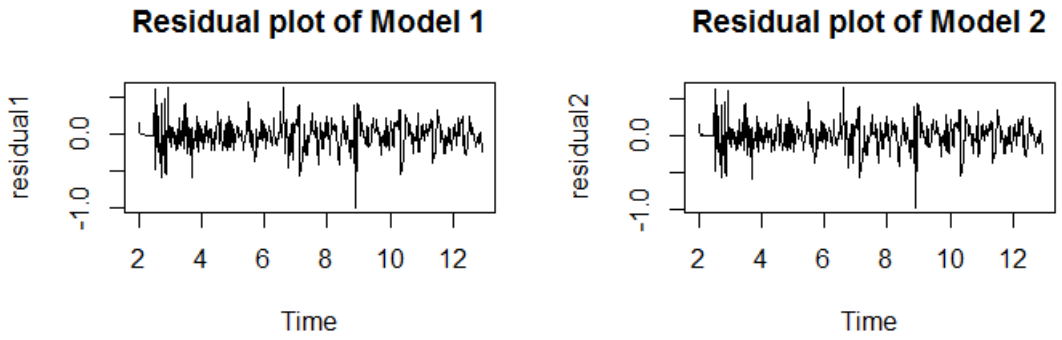
For Holt-Winters, we use the additive and multiplicative filtering methods. When looking through both methods, we see that both methods produce almost the same results. We can see this clearly by looking at the additive filtering and multiplicative filtering plots shown below in figure 3.1.1. Thus, we will consider both models for our candidate model.

Figure 3.1.1: Plots for Holt Winters Additive and Multiplicative Filtering



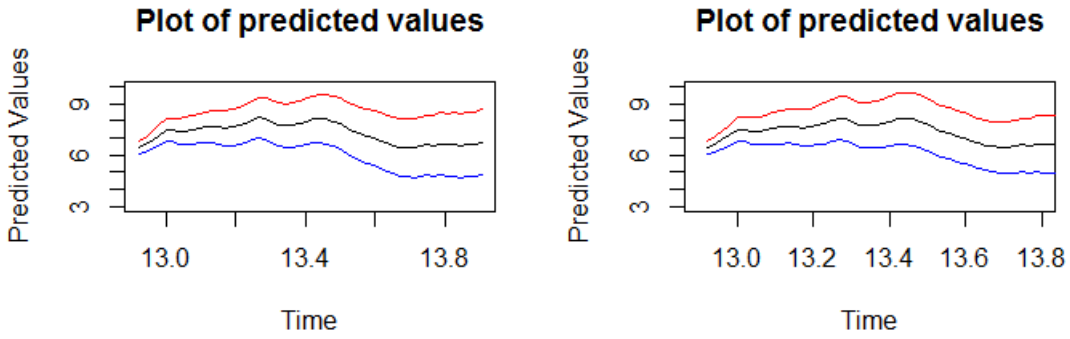
First, we look at the residual plots to see if the models do not have a distinct pattern. By looking at the residual plots (figure 3.1.2), both plots look almost identical. The good thing that we notice from the residual plots was that there is no distinct pattern in the residual plots. Thus, we can continue forward with the smoothing.

Figure 3.1.2: Residual Plots for Model 1 and 2



For both methods we did a 95 % prediction interval. We plotted the predictions for each time point. We also included the upper bound and lower bound of the predictions, which is our prediction interval.

Figure 3.1.3: Plots of Predicted Values for Model 1 and 2



To test the accuracy of the model, we used the Shapiro-Wilk test.

- For model 1, the Shapiro-Wilk test is 0.96936 with a p-value of 1.628e-09
- For model 2, the Shapiro-Wilk test produced 0.96909 and a p-value of 1.435e-09

Since both methods resemble each other and both have fairly close Shapiro-Wilk test values, with the smaller one being Model 2, we will consider them both in our final comparison for a model for our data. Simultaneously, it can be noted that Model 2 may be advantageous when compared to model 1.

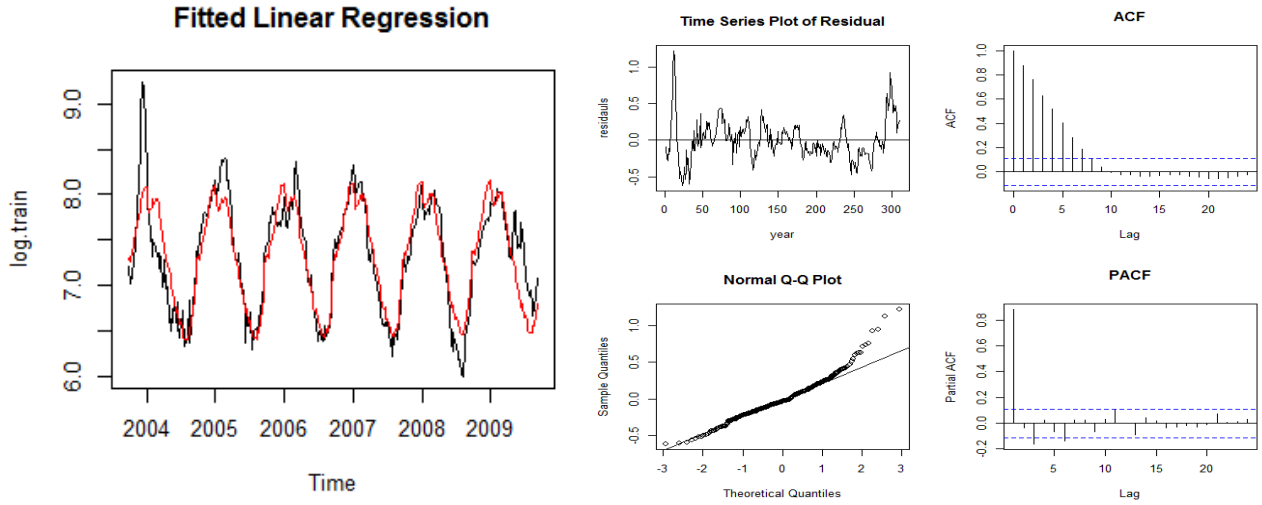
3.2 Regression Analysis

3.2.1 Model Selection

Model 1 – Log Additive Linear Model

The simplest model would be an additive linear model between time and week number. Those two variates are fitted to the logarithmic flu trend of Ontario. From the result below, we see that the fitted data, in general, poorly represents the observed data. At around Time 2004 and 2008.6 we can notice that there are evident outliers. Additionally, the normality assumption is violated as the QQ-plot is heavily tailed. The ACF plot is also extremely volatile.

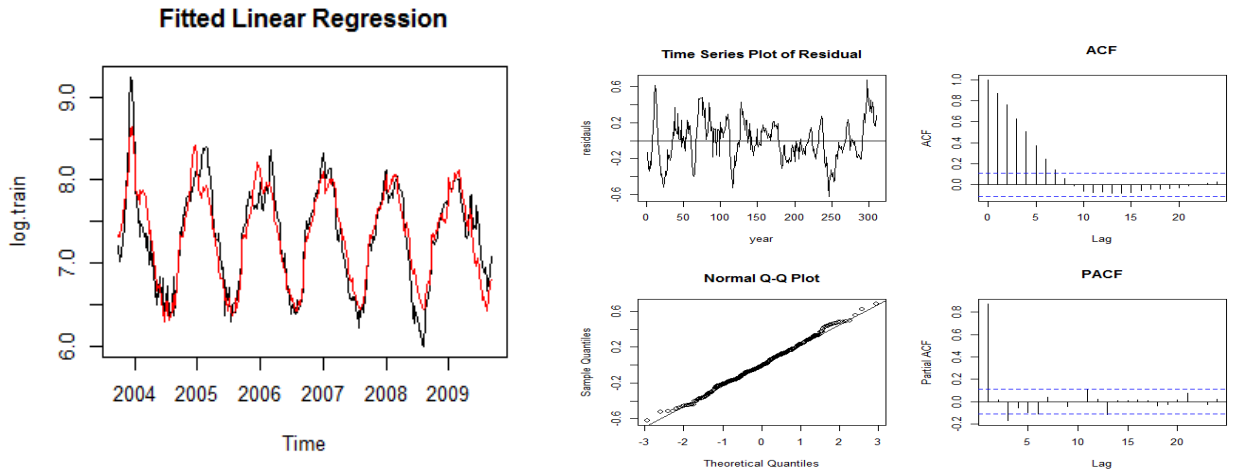
Figure 3.2.1: Fitted Linear Regression and Diagnostic Graphs for Model 1



Model 2 – Log Full Interactive Linear Model

The second model would be an interactive model between time and week number. Once again, a full interaction model of these two variates are fitted to the logarithmic flu trend of Ontario. Much like model 1, the fitted data poorly represents the observed data. Outliers are present in similar positions. Normality is much improved in this version as evident by the QQ plot. However, the ACF remains spiked.

Figure 3.2.2: Fitted Linear Regression and Diagnostic Graphs for Model 2

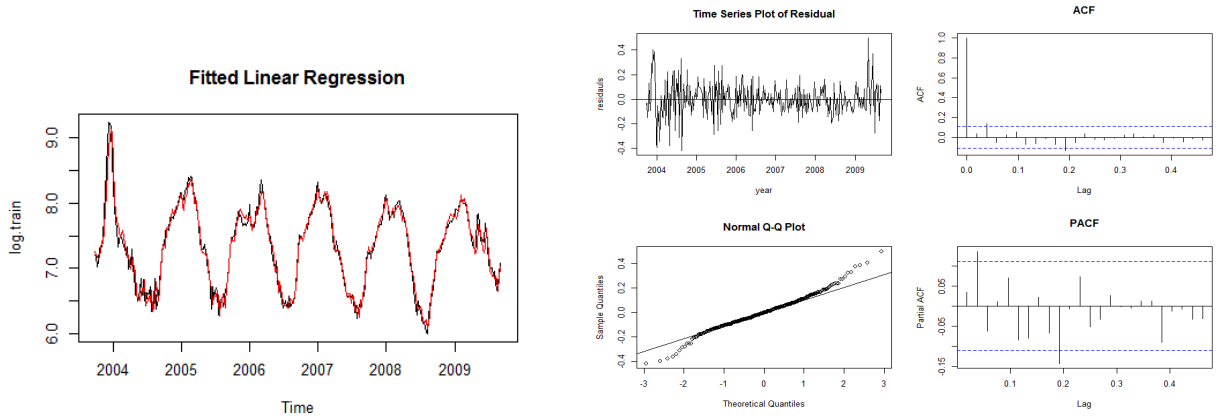


From the analysis of the two linear plots above, we can conclude that linear regression cannot accurately model the observed data. Model 1 and Model 2 will not be chosen for forecasting. Additionally, when analyzing the ACF plot, despite its spikes we can see a pattern resembling exponential decay. Further analysis yields 4 potential results: ARMA(1,0), ARMA(1,1), ARMA(2,1), ARMA(2,2). Below are the residual analyses of the four potential models.

Model 3 – ARMA (1,0) of Additive Model

From graphical analysis, the fitted values are very close to the observed values. The outliers found in the previous two points have been accounted for. Normality is much better than the additive model but is worse than the multiplicative model. This is explainable as the multiplicative model was severely over-fitted. The spiked problems appear to be correct, although a few jumps still exist at lag 2 and 10.

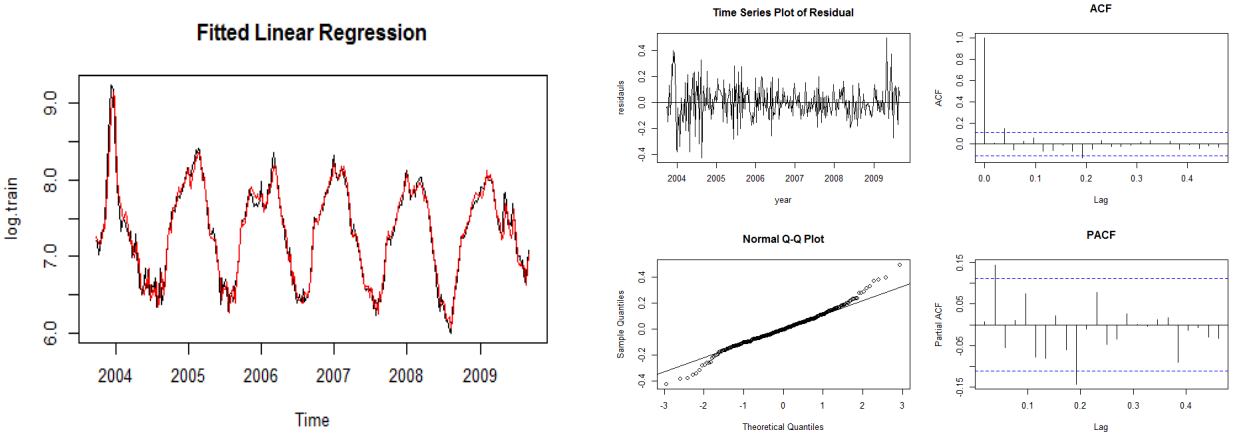
Figure 3.2.3: Fitted Linear Regression and Diagnostic Graphs for Model 3



Model 4 – ARMA (1,1) of Additive Model

From graphical analysis, this model is identical to Model 3: ARMA(1,0).

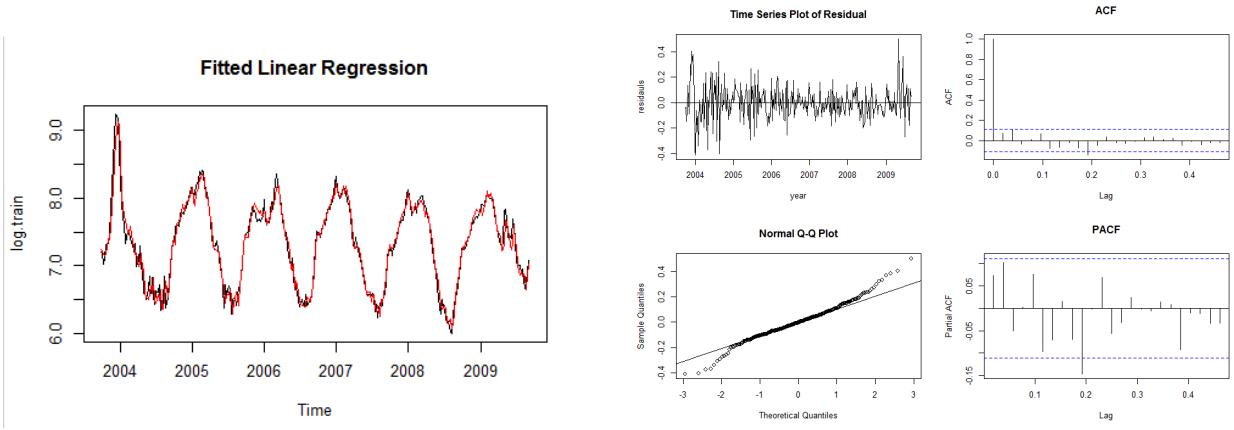
Figure 3.2.4: Fitted Linear Regression and Diagnostic Graphs for Model 4



Model 5 – ARMA (2,1) of Additive Model

Much like the previous two models, the observed values are very close. Notice that the jump at lag 2 is gone and that the PACF values are much more fitted to 0.

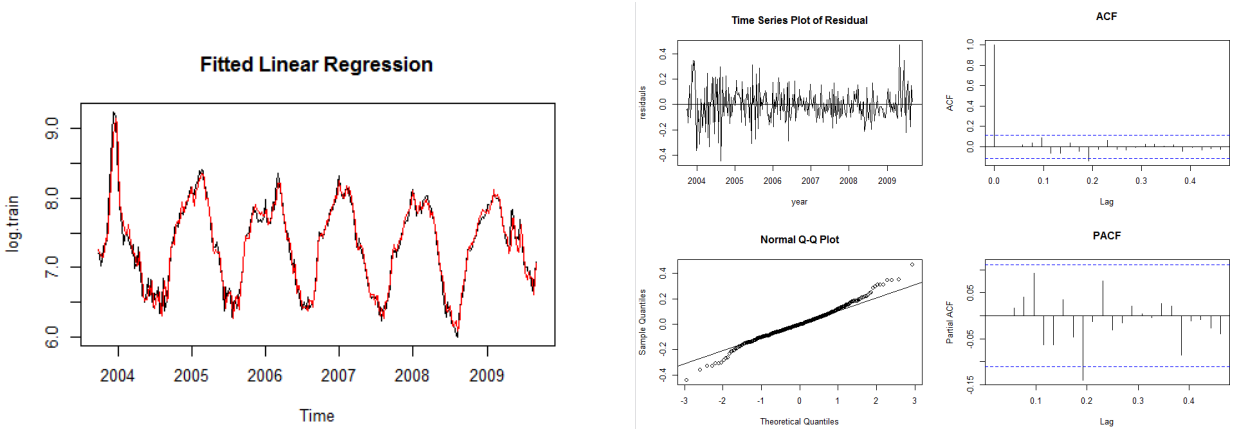
Figure 3.2.5: Fitted Linear Regression and Diagnostic Graphs for Model 5



Model 6 – ARMA (2,2) of Additive Model

This model is very similar to the previous model in terms of graphical analysis. There is a slight improvement in the QQ plot normality as well as the jumps in the ACF and PACF plot.

Figure 3.2.6: Fitted Linear Regression and Diagnostic Graphs for Model 6



3.2.2 Model Analysis

```
       Model3    Model4    Model5   Model6
AIC -283.4902 -281.7923 -280.4735 -286.765
```

```
        Model3  Model4   Model5   Model6
PRESS 5.076144 5.07116 5.059886 6.590558
```

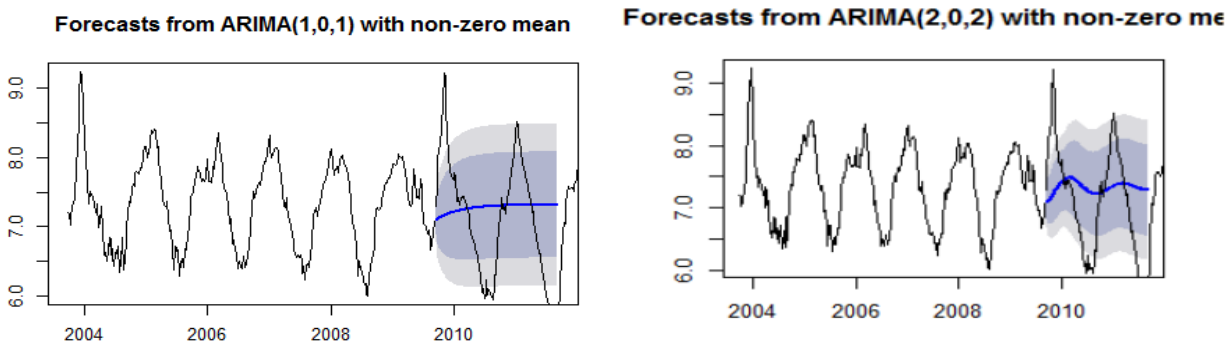
When comparing ARMA(1,1) to ARMA(1,0) the result has the same residual diagnostics. However, notice that the press is reduced. Therefore model 4 would be the superior model. When removing the remaining jumps in the ACF and PACF plot, we ran into the ARMA(2,1) and ARMA(2,2) model. ARMA(2,2) completely removes all the spikes as well as yields a better AIC value. Hence Model 6 is a clear candidate model as well.

"From the analysis of both the previous graphics, the AIC, and the PRESS residuals, our group suggests that model 4 and model 6 would be the optimal models chosen based on regression analysis."

3.2.3 Model Forecasting

From the two chosen models, the graphs below show the 80%, 95% and the direct forecasting of our testing set.

Figure 3.2.7: ARIMA(1,0,1) and ARIMA(2,0,2) Forecasts



3.3 Box-Jenkins Models

After examining the ACF and PACF graphs (figure 3.3.1) of the original data and the log transformed data, the graphs appear to follow a sin-like function. Various combinations of seasonality and ordinarity were attempted to create a stationary process. The data with one difference in seasonality and ordinarity is the closest to a stationary process. Plotting the ACF and PACF graph (figures 3.3.2-3.3.3) shows exponential decay. We suggest the following possible models for both the original data and the log transformed data:

- SARIMA (2,1,2)(0,1,1)
- SARIMA (2,1,1) (0,1,1)
- SARIMA (1,1,1) (0,1,1)
- SARIMA(1,1,2) (0,1,1)

Figure 3.3.1: ACF and PACF plots for the ordinary training set and log training set

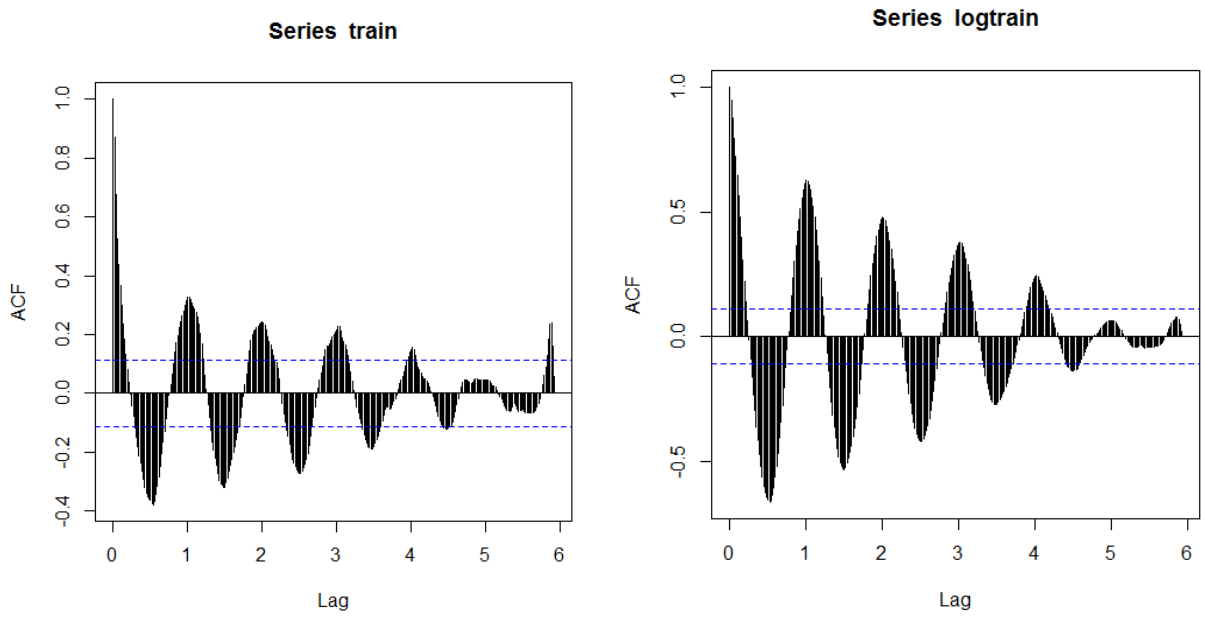


Figure 3.3.2: ACF and PACF plots for the data with one difference

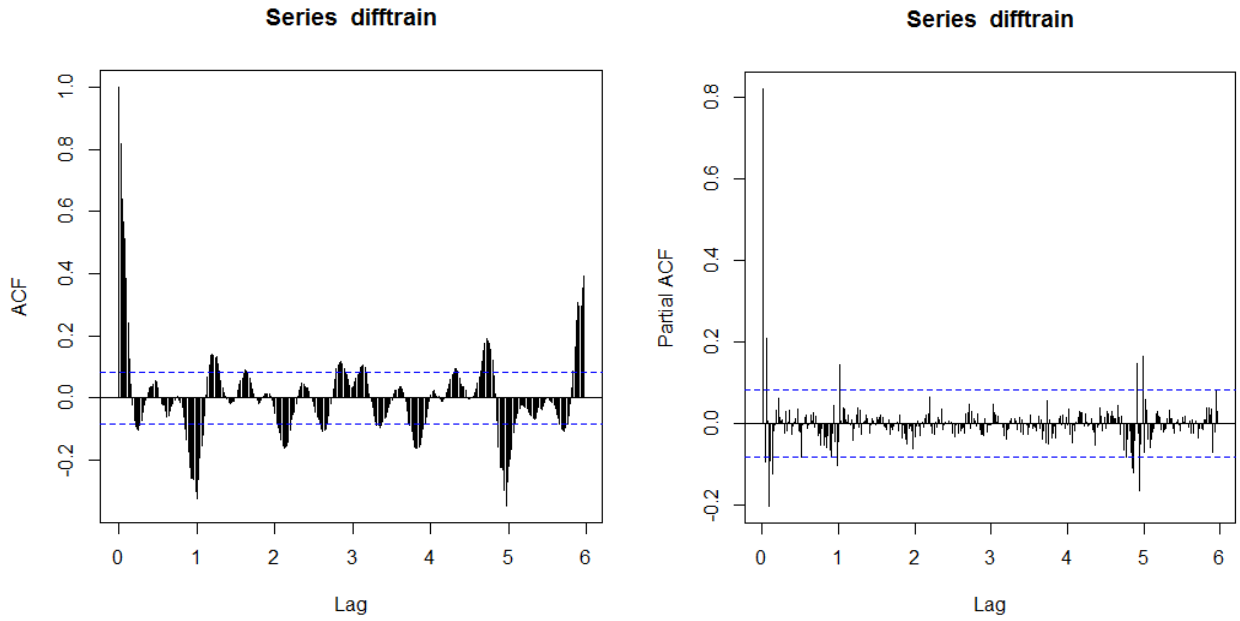
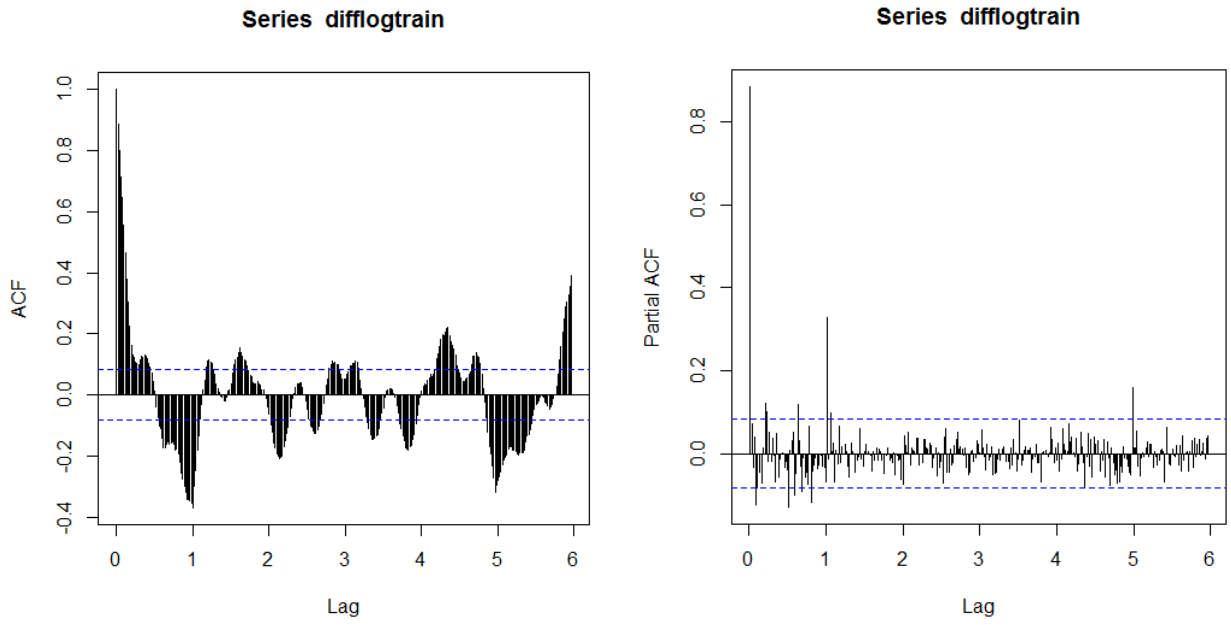


Figure 3.3.3: ACF and PACF plots for the log of the data with one difference



The original data is fitted with the four potential models and the PRESS and AIC values are shown in Table 3.3.1. Model 1 outperforms with the smallest AIC while Model 4 outperforms with the smallest PRESS.

Table 3.3.1: Comparison between 4 Models with Original Data

	SARIMA (2,1,2)(0,1,1)	**SARIMA (2,1,1)(0,1,1)**	**SARIMA (1,1,1)(0,1,1)**	**SARIMA (1,1,2)(0,1,1)**
PRESS	3033437349	3624151583	2060589902	869736407
AIC	3899.665	3909.311	3916.082	3914.507

The log transformed data is fitted with the four potential models and the PRESS and AIC values are shown in Table 3.3.2. Model 3 has the smallest PRESS and AIC value so it outperforms the other 3 models.

Table 3.3.2: Comparison between 4 Models with Log Transformed Data

	SARIMA (2,1,2)(0,1,1)	**SARIMA (2,1,1)(0,1,1)**	**SARIMA (1,1,1)(0,1,1)**	**SARIMA (1,1,2)(0,1,1)**
PRESS	2331461430	2313577185	812949481	862593613
AIC	-161.1932	-161.9326	-164.7747	-162.8304

Comparing Table 3.3.1 and Table 3.3.2, it appears that using the log transformed data provides a better fit than the original data. Predicted value plots, QQ-plots, ACF and PACF (Figures 3.3.4-3.3.7) were only created for the log transformed data since it appears to be the better fit. None of the models' prediction intervals contain all the actual values, the first few weeks of actual values are not within the prediction interval.

Comparing the actual values to the predicted values, Model 3 does the best job of predicting the seasonality of the actual data. The ACF and PACF of all the residuals appear to be very similar with spikes every now and then.

"Overall, Model 3 is the best candidate model for Box-Jenkins method."

Figure 3.3.4: Model 1 Diagnostic Plots

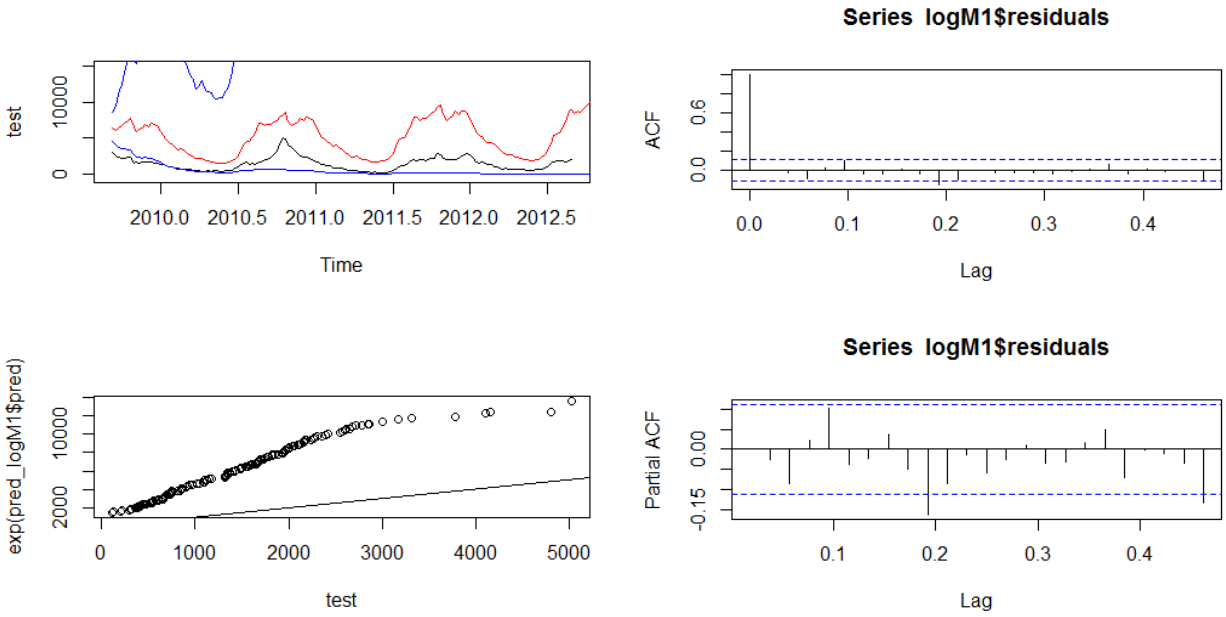


Figure 3.3.5: Model 2 Diagnostic Plots

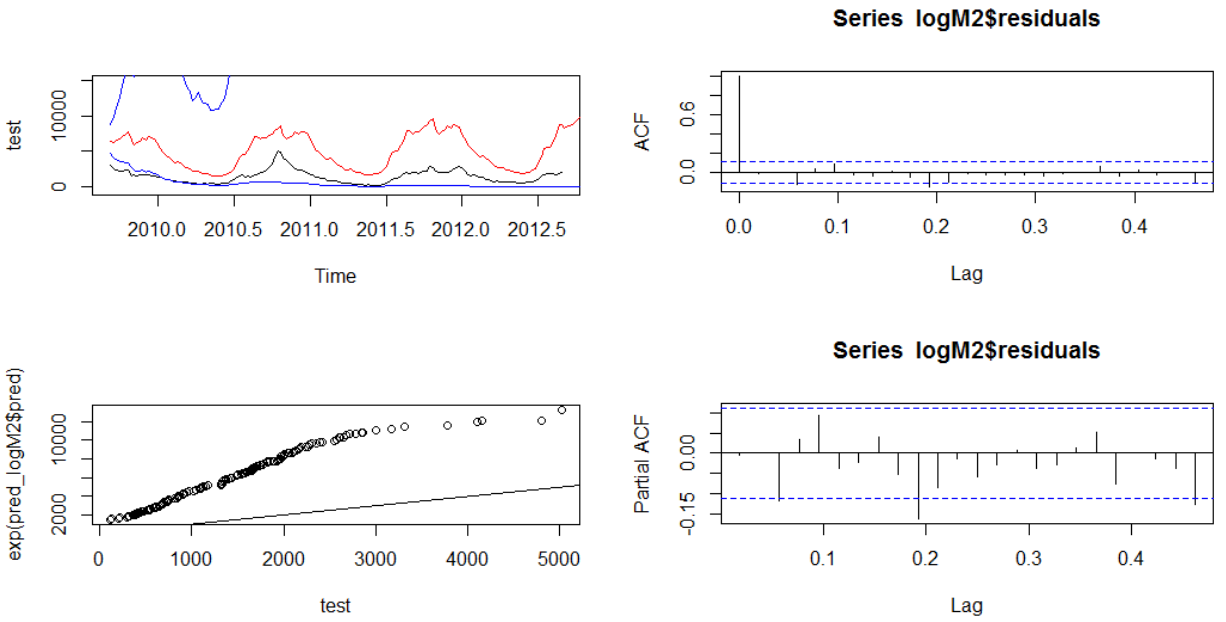


Figure 3.3.6: Model 3 Diagnostic Plots

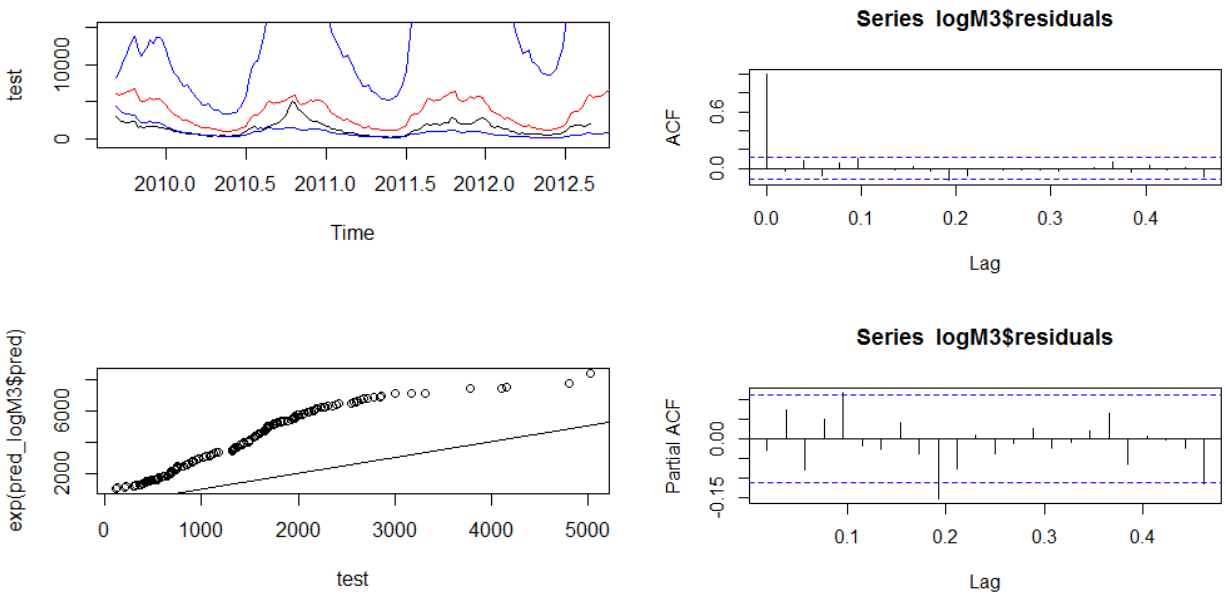
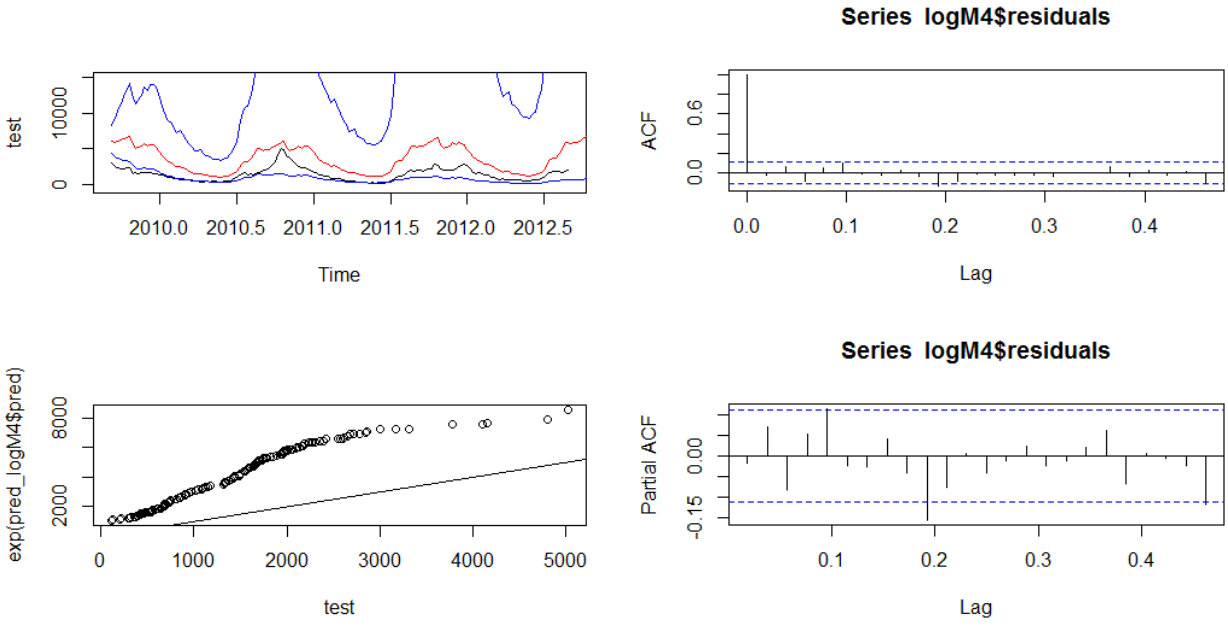


Figure 3.3.7: Model 4 Diagnostic Plots



4.0 Conclusions

Based on the above Data Analysis, the best models for each of the targeted Statistical Analyses are listed below, named Models 1 to 5 for the sake of comparison. We will now do a final analysis of these models in comparison with each other in order to find the model that is best suited for our data.

Holt Winters: Model 1: Holt-Winters Additive Model

Model 2: Holt-Winters Multiplicative Model

Regression: Model 3: ARMA(1,1) of the Additive Model

Model 4: ARMA(2,2) of the Additive Model

Box-Jenkins: Model 5: SARIMA(1,1,1)(0,1,1)

4.1 Statistical Conclusions

The optimal models mentioned above were compared using Press Residual and AIC analyses as well as an interpretation of their plots. The results are as follows:

Table 4.1.1: Table of Press Residuals and AIC values for the 5 optimal models

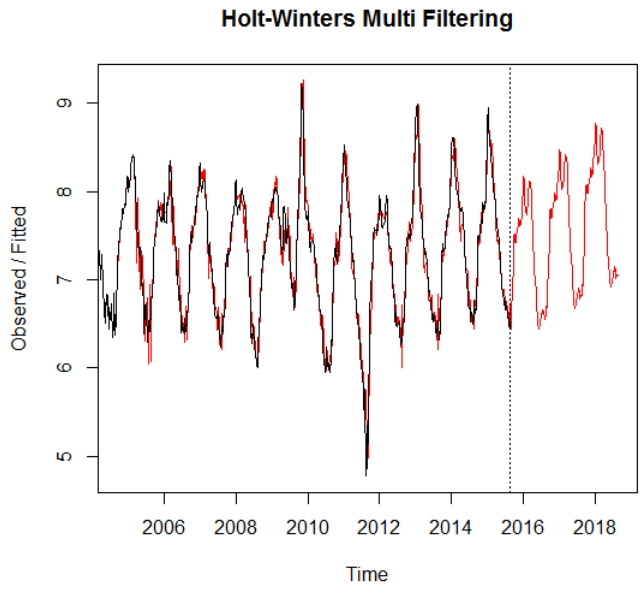
	Model 1	**Model 2**	**Model 3**	**Model 4**	**Model 5**
PRESS	3.1094	2.9652	5.0712	6.5906	2060589902
AIC	N/A	N/A	-287.7923	-286.7650	3916.0820

Evidently, our best model, based on the above criteria is Model 2 as this model has the PRESS Residual. Undergoing AIC analysis, it can be observed that Model 4 has the smallest AIC which is desirable but since AIC cannot be applied to the Holt-Winters models, PRESS statistics are a better choice for comparison.

For our graphical analysis, we will observe the plots for all 5 optimal models: figures 3.1.3, 3.2.7, 3.3.6. These plots show that all of the optimal models in some way, fit the data well. However, Model 2 seems to fit the best, agreeing with our PRESS statistic analysis.

"Therefore, we will use Model 2 to forecast our data."

Figure 4.1.1: 155-week forecast for Optimal Model 2: Holt-Winters Multiplicative model



4.2 Interpretations and Limitations

From the forecast shown in figure 4.1.1, a increasing trend is noticed within the forecast period. The graph also shows strong seasonality, proving that it was a good choice of a model as the original data also underwent seasonal effects. Particularly, the increasing trend is a cause for concern as we hypothesized a decrease in flu records. That being said, we fail to accept our null hypothesis.

Even though the rise in level is slow, it is shown over a period of 155 weeks and for just one Province. If the prediction turns out to be accurate, this could lead to off- the- charts flu records for years after that. Therefore, there is a need for future research on the matter with more variables to be considered so that prevention procedures can be put in place to ease the increase in flu trends.

As with all Statistical Analyses, there were some limitations to our project that may reshape the actual forecast in flu trends. Due to missing data in the time series, our analysis may not be as accurate as it should be. One recommendation for future analysis is to use some method to impute missing data, for example, an Expectation-Maximization Algorithm.

Despite the limitation in the scope of our Statistical Methods, the Analysis was successful in choosing an optimal model to forecast the future.

5.0 References

Data

https://www.google.org/flutrends/about/data/flu/ca/data.txt

Course Notes

"Stat 443: Forecasting", Paul Marriott, January 3rd, 2017.

6.0 Appendix: R code

```
#Data Analysis ********************************************************************************

flu.data <- read.table("C:/Users/sraml/Dropbox/waterloo/year4/4B/stat443/project/data.txt", header=TRUE, sep=",", skip
= 11)
ont.data <- flu.data[,c("Ontario")]
flu.ts <- ts(ont.data, start = c(2003,39), frequency = 52)

train <- ts(ont.data[1:310], start = c(2003, 39), frequency = 52)   #50% training set
test <- ts(ont.data[311:465], start = c(2009, 37), frequency = 52)  #25% testing set
valid <- ts(ont.data[466:620], start = c(2012, 36), frequency = 52) #25% validating set

# Log
log.train <- log(train)
log.test <- log(test)
log.valid <- log(valid)

# Plot data
plot(flu.ts, main="Ontario Flu Trend")

# Plot decomposed data
plot(decompose(flu.ts))

# Compute and Plot the logarithm of the data
logflu<-log(flu.ts)
plot(logflu, main="Logarithm of the Ontario Flu Data")

# Useful Functions
install.packages("forecast")
library(forecast)

PRESS <- function(prediction) {
  #' calculate the PRESS
  PRESS <- sum(prediction$residual^2)
  return(PRESS)
}

analyze <- function(value) # Residual Analysis
{
  par(mfcol=c(2,2)) # Compact the aforementioned plots
  ts.plot(value, gpars=list(xlab="year", ylab="residauls")) # 1st plot: TS of residuals
  title("Time Series Plot of Residual")
  abline(h=mean(value)) # Mean 0 Line

  qqnorm(value) # 2nd plot: QQplot
  qqline(value)

  ACF <- acf(value, main ="ACF") # 3rd plot: ACF
  PACF <- acf(value, main ="PACF", type="partial") # 4th plot: PACF
}

#Holt Winters ********************************************************************************

# Plot the time series
```

```
plot(flu.ts, ylab = "Flu", main = "Time series data of the Flu")
# Seasonal effect is there but the variance is too large

plot(logflu, ylab = "Log of Flu Data", main = "Time Series of Log Flu Data")
# Variance is reduced

# Defining data for Holt Winters Additive model
model1 = HoltWinters(logflu, seasonal="add")
#Plotting Holt Winters additive model with a 155 week predition
plot(model1, predict(model1, n.ahead = 155), main = "Holt-Winters Additive Filtering")
residual1 = residuals(model1)
plot(residual1, main = "Residual plot of Model 1")
Press_model1 = sum((test-exp(predict(model1, n.ahead = 155, interval = "pred")$pred[1:155]))^2)

# Defining data for Holt Winters Multiplicative model
model2 = HoltWinters(logflu, seasonal="mult")
#Plotting Holt Winters multiplicative model with a 155 week predition
plot(model2, predict(model2, n.ahead = 155), main = "Holt-Winters Multi Filtering")
residual2 = residuals(model2)
plot(residual2, main = "Residual plot of Model 2")
Press_model2 = sum((test-exp(predict(model2, n.ahead = 155, interval = "pred")$pred[1:155]))^2)

# 95% Prediction interval for models 1 and 2
# prediction interval for the Holt-Winter transformation of the data
PIadd = predict(model1, 155, prediction.interval = TRUE)
# plotting the prediction values for the next 52 weeks
plot(PIadd[,1], ylim = c(3,10), ylab = "Predicted Values",main = "Plot of predicted values")
lines(PIadd[,2], col="red") #Plotting upper bound for the prediction interval
lines(PIadd[,3], col="blue") #Plotting lower bound for the prediction interval
summary(model1)

# prediction interval for the Holt-Winter transformation of the data
PImult = predict(model2, 155, prediction.interval = TRUE)
# plotting the prediction values for the next 52 weeks
plot(PImult[,1], xlim = c(12.9, 13.8),ylim = c(3,10), ylab = "Predicted Values",main = "Plot of predicted values")
lines(PImult[,2], col="red") #Plotting upper bound for the prediction interval
lines(PImult[,3], col="blue") #Plotting lower bound for the prediction interval
summary(model2)

# Shapiro-Wilk Test
residual1 = residuals(model1)
plot(residual1)
residual2 = residuals(model2)
plot(residual2)
shapiro.test(residual1)# W = 0.96936, p-value = 1.628e-09
shapiro.test(residual2)# W = 0.96909, p-value = 1.435e-09


#Regression ********************************************************************************

# Training Set
time <- time(log.train)
week <- as.factor(cycle(log.train))

# Testing set
time.test<-time(log.test)
week.test<-as.factor(cycle(log.test))
fit.test<-lm(log.test~ time.test+ week.test)
```

```
X2 <-model.matrix(fit.test)

## Model Selection
# Model 1: Additive Model
model1_fit <- lm(log.train ~ time + week)
plot(log.train, main = "Fitted Linear Regression")
points(time, model1_fit$fitted, type='l', col="red")
residual <- residuals(model1_fit)
analyze(residual)
summary(model1_fit)

# Model 2: Interaction Model
model2_fit <- lm(log.train ~ (time + week)^2)
plot(log.train, main = "Fitted Linear Regression")
points(time, model2_fit$fitted, type='l', col="red")
residual <- residuals(model2_fit)
analyze(residual)

# Analysis yields that ARMA model is better predictor
## Start with ARMA (1,0)

# Model 3: ARMA(1,0) of additive model
X <- model.matrix(model1_fit)
model3_arma <-arima(log.train, order = c(1,0,0), xreg = X[,2:53]) # 2:53 because weeks
plot(log.train, main = "Fitted Linear Regression")
points(time, log.train - model3_arma$res, type='l', col="red")
analyze(residuals(model3_arma))

# Model 4: ARMA(1,1) of additive model (To handle unexplained spike at lag 1 and 2)
X <- model.matrix(model1_fit)
model4_arma <-arima(log.train, order = c(1,0,1), xreg = X[,2:53])
plot(log.train, main = "Fitted Linear Regression")
points(time, log.train - model4_arma$res, type='l', col="red")
analyze(residuals(model4_arma))

# Model 5: ARMA(2,1) of additive model (one spiked reduced, one spike remains)
X <- model.matrix(model1_fit)
model5_arma <-arima(log.train, order = c(2,0,1), xreg = X[,2:53])
plot(log.train, main = "Fitted Linear Regression")
points(time, log.train - model5_arma$res, type='l', col="red")
analyze(residuals(model5_arma))

# Model 6: ARMA(2,2) of additive model
X <- model.matrix(model1_fit)
model6_arma <-arima(log.train, order = c(2,0,2), xreg = X[,2:53])
plot(log.train, main = "Fitted Linear Regression")
points(time, log.train - model6_arma$res, type='l', col="red")
analyze(residuals(model6_arma))

# Prediction Model 4 (Best Model with ARMA with p = 1)
model4_pred <- forecast(model4_arma, n.ahead = 155, xreg = X2[,2:53]) # prediction for 155 weeks
plot(model4_pred)
lines(log.test)

# Prediction Model 6 (Best Model with ARMA with p = 2)
model6_pred <- forecast(model6_arma, n.ahead = 155, xreg = X2[,2:53]) # prediction for 155 weeks
plot(model6_pred)
lines(log.test)
```

```
# Analysis
model3_pred <- forecast(model3_arma, n.ahead = 155, xreg = X2[,2:53])
model5_pred <- forecast(model5_arma, n.ahead = 155, xreg = X2[,2:53])

AIC <- cbind(Model3 = model3_arma$aic, Model4 = model4_arma$aic, Model5 = model5_arma$aic, Model6 =
model6_arma$aic)
rownames(AIC) <- "AIC"
press <- cbind(Model3 = PRESS(model3_pred), Model4 = PRESS(model4_pred), Model5 =PRESS(model5_pred),
Model6 =PRESS(model6_pred))
rownames(press) <- "PRESS"


#Box-Jenkins ********************************************************************************

par(mfrow=c(1,1))

difftrain <- diff(train,lag = 52)
difflogtrain <- diff(log.train,lag = 52)
acf(train, lag.max = 310)
acf(train, lag.max = 310, type ="partial")
acf(difftrain, lag.max = 310)
acf(difftrain, lag.max = 310, type ="partial")
acf(log.train, lag.max = 310)
acf(difflogtrain, lag.max = 310)
acf(difflogtrain, lag.max = 310, type = "partial")

plot.ts(train)
plot.ts(log.train)
plot.ts(difftrain)
plot.ts(diff(difftrain))
plot.ts(difflogtrain)
plot.ts(diff(difflogtrain))

M1 <- arima(train,order=c(2,1,2), seasonal=list(order=c(0,1,1), period=52))
M2 <- arima(train,order=c(2,1,1), seasonal=list(order=c(0,1,1), period=52))
M3 <- arima(train,order=c(1,1,1), seasonal=list(order=c(0,1,1), period=52))
M4 <- arima(train,order=c(1,1,2), seasonal=list(order=c(0,1,1), period=52))

pred_M1 = predict(M1,n.ahead = 310, interval = "pred")
Press_M1 = sum((test-pred_M1$pred[1:155])^2)
plot.ts(test, ylim= c(-500, 15000))
points(pred_M1$pred , type = "l", col = "red")
points(pred_M1$pred + 1.96*pred_M1$se, type= "l", col = "blue")
points(pred_M1$pred - 1.96*pred_M1$se, type= "l", col = "blue")
qqplot(test,pred_M1$pred)

pred_M2 = predict(M2,n.ahead = 310, interval = "pred")
Press_M2 = sum((test-pred_M2$pred[1:155])^2)
plot.ts(test, ylim= c(-500, 15000))
points(pred_M2$pred , type = "l", col = "red")
points(pred_M2$pred + 1.96*pred_M2$se, type= "l", col = "blue")
points(pred_M2$pred - 1.96*pred_M2$se, type= "l", col = "blue")
qqplot(test,pred_M2$pred)

pred_M3 = predict(M3,n.ahead = 310, interval = "pred")
Press_M3 = sum((test-pred_M3$pred[1:155])^2)
plot.ts(test, ylim= c(-500, 15000))
```

```
points(pred_M3$pred , type = "l", col = "red")
points(pred_M3$pred + 1.96*pred_M3$se, type= "l", col = "blue")
points(pred_M3$pred - 1.96*pred_M3$se, type= "l", col = "blue")
qqplot(test,pred_M3$pred)

pred_M4 = predict(M4,n.ahead = 310, interval = "pred")
Press_M4 = sum((test-pred_M4$pred[1:155])^2)
plot.ts(test, ylim= c(-500, 15000))
points(pred_M4$pred , type = "l", col = "red")
points(pred_M4$pred + 1.96*pred_M4$se, type= "l", col = "blue")
points(pred_M4$pred - 1.96*pred_M4$se, type= "l", col = "blue")
qqplot(test,pred_M4$pred)

logM1 <- arima(log.train,order=c(2,1,2), seasonal=list(order=c(0,1,1), period=52))
logM2 <- arima(log.train,order=c(2,1,1), seasonal=list(order=c(0,1,1), period=52))
logM3 <- arima(log.train,order=c(1,1,1), seasonal=list(order=c(0,1,1), period=52))
logM4 <- arima(log.train,order=c(1,1,2), seasonal=list(order=c(0,1,1), period=52))

pred_logM1 = predict(logM1,n.ahead = 310, interval = "pred")
Press_logM1 = sum((test-exp(pred_logM1$pred[1:155]))^2)
plot.ts(test, ylim= c(-500, 15000))
points(exp(pred_logM1$pred) , type = "l", col = "red")
points(exp(pred_logM1$pred + 1.96*pred_logM1$se), type= "l", col = "blue")
points(exp(pred_logM1$pred - 1.96*pred_logM1$se), type= "l", col = "blue")
qqplot(test,exp(pred_logM1$pred))
abline(1,1)
acf(logM1$residuals)
acf(logM1$residuals, type = "partial")

pred_logM2 = predict(logM2,n.ahead = 310, interval = "pred")
Press_logM2 = sum((test-exp(pred_logM2$pred[1:155]))^2)
plot.ts(test, ylim= c(-500, 15000))
points(exp(pred_logM2$pred) , type = "l", col = "red")
points(exp(pred_logM2$pred + 1.96*pred_logM2$se), type= "l", col = "blue")
points(exp(pred_logM2$pred - 1.96*pred_logM2$se), type= "l", col = "blue")
qqplot(test,exp(pred_logM2$pred))
abline(1,1)
acf(logM2$residuals)
acf(logM2$residuals, type = "partial")


pred_logM3 = predict(logM3,n.ahead = 310, interval = "pred")
Press_logM3 = sum((test-exp(pred_logM3$pred[1:155]))^2)
plot.ts(test, ylim= c(-500, 15000))
points(exp(pred_logM3$pred) , type = "l", col = "red")
points(exp(pred_logM3$pred + 1.96*pred_logM3$se), type= "l", col = "blue")
points(exp(pred_logM3$pred - 1.96*pred_logM3$se), type= "l", col = "blue")
qqplot(test,exp(pred_logM3$pred))
abline(1,1)
acf(logM3$residuals)
acf(logM3$residuals, type = "partial")


pred_logM4 = predict(logM4,n.ahead = 310, interval = "pred")
Press_logM4 = sum((test-exp(pred_logM4$pred[1:155]))^2)
plot.ts(test, ylim= c(-500, 15000))
points(exp(pred_logM4$pred) , type = "l", col = "red")
points(exp(pred_logM4$pred + 1.96*pred_logM4$se), type= "l", col = "blue")
```

```
points(exp(pred_logM4$pred - 1.96*pred_logM4$se), type= "l", col = "blue")
qqplot(test,exp(pred_logM4$pred))
abline(1,1)
acf(logM4$residuals)
acf(logM4$residuals, type = "partial")

AIC(M1)
AIC(M2)
AIC(M3)
AIC(M4)
Press_M1
Press_M2
Press_M3
Press_M4
min(AIC(M1),AIC(M2),AIC(M3),AIC(M4))
min(Press_M1,Press_M2,Press_M3,Press_M4)

AIC(logM1)
AIC(logM2)
AIC(logM3)
AIC(logM4)
Press_logM1
Press_logM2
Press_logM3
Press_logM4
min(AIC(logM1),AIC(logM2),AIC(logM3),AIC(logM4))
min(Press_logM1,Press_logM2,Press_logM3,Press_logM4)
shapiro.test(logM1$residuals)
```