DATA 624: Project 1

Bethany Poulin

October 22, 2019

Contents

Overview

I am leaving the project overview page here for us to compile our final report in one singular document. We will add additional information here regarding project one to include explanation of process, etc.

Dependencies

Please add all libraries used here.

The following R libraries were used to complete Project 1:

```r
# General
library('easypackages')

libraries('knitr', 'kableExtra', 'default')

# Processing
libraries('readxl', 'tidyverse', 'janitor', 'lubridate')

# Graphing
libraries('ggplot2', 'grid', 'gridExtra', 'ggfortify','ggpubr')

# Timeseries
libraries('zoo', 'urca', 'tseries', 'timetk')

# Math
libraries('forecast')
```

Data

Data was stored within our group repository and imported below using the `readxl` package. Each individual question was solved within an R script and the data was sourced into our main report for discussion purposes. The R scripts are available within our appendix for replication purposes.

For grading purposes, we exported and saved all forecasts as a csv in our data folder.

```r
# Data Aquisition
waterflow_1 <- read_excel("data/Waterflow_Pipe1.xlsx")
waterflow_2 <- read_excel("data/Waterflow_Pipe2.xlsx")

# Source Code
source("scripts/Part-C-BP.R")
```

1 Part C

Part C.consists of two data sets. These are simple 2 columns sets, however they have different time stamps. Your assignment is to time-base sequence the data and aggregate based on hour (example of what this looks like, follows). Note for multiple recordings within an hour, take the mean. Then to test appropriate assumptions and forecast a week forward with confidence bands (80 and 95%). Add these to your existing files above – clearly labeled.

1.1 Exploration

Pipe one:
* 1000 observations
* No missing values
* Multiple reading within each hour
* 9-days of data

Pipe Two
* 100 Observations
* No missing values
* Single reading on the hour
* 41-days of data

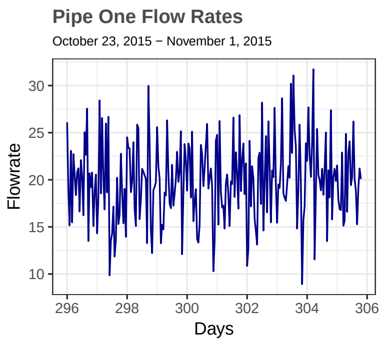
Because of the disparities in the data some grooming was necessary.
For Pipe One, representing 9-days of water flow rate measurements multiple samples per hour, a mean of all rates in the hour was taken and labeled with the whole-hour at the beggining of the period (floor hour) to align with the hourly readings from Pipe Two.

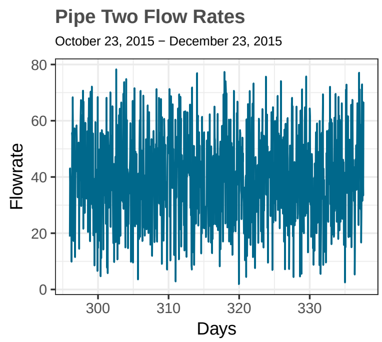
After aggregating, there were only 236 observations (spanning 9-days) of pipe one and still 1000 observations (spanning 41-days) from Pipe Two.

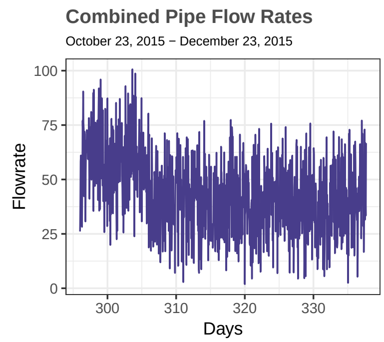
These data posed an interesting conundrum. With two possible ways of handling it.
- Merge the files, and use only 236 observations
- all forecasts would be based on the combined data
- this would mean making 168 forecasts with only 236 data-points prior
- all forecasts would be starting November 1, instead of from the end of data December 3
- Merge the files and use the whole set to make predictions
- we would have 100 observations to model prior to forecasts
- 236 of the observations would be be different from the remaining 764, which could both alter the model type and forecast
- we would be forecasting from the natural ending of tPipe Two readings

Because it was concievable that there might be a daily periodicity, it was important to have a frequency of 24, which made numbering by day of year and grooming the time series to start on the 7081 hour aligning with October 23 01:00 AM.

1.1.1 Time Series Plots

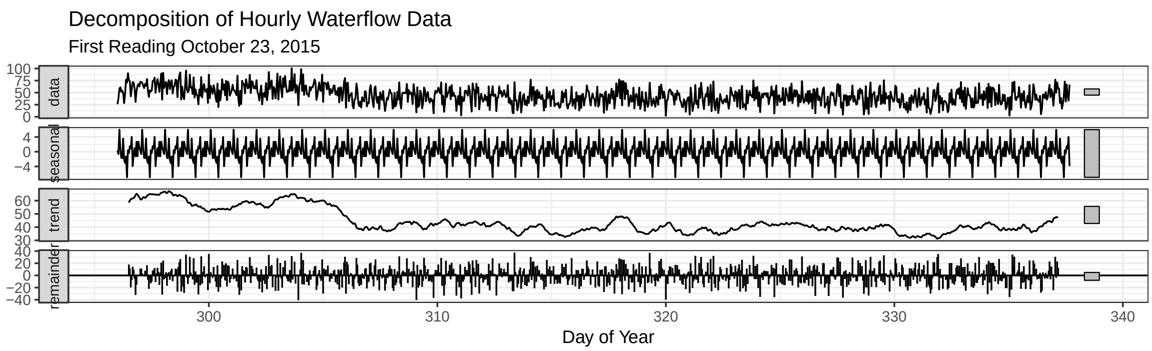






1.1.2 Decomposition

It is clear from the combined plot that there is a pretty notable change in the trend when the readings from Pipe One wane. Let's look at the decomposed seriesand see if it gives us some insight into a good model.



From the decomposition, the appears to be a seasonal component in agreement with the assessment that there might be a daily flowrate periodicity. Also, as expected, around day 306 where Pipe One flow rates go silent there is a trend down and then relatively flat trend thereafter.

1.2 Estimating Stationarity

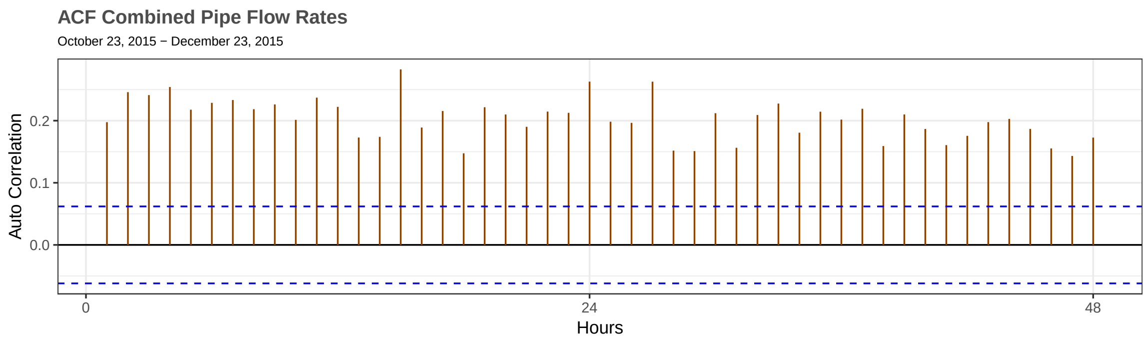
Number of Estimated Differences: 1

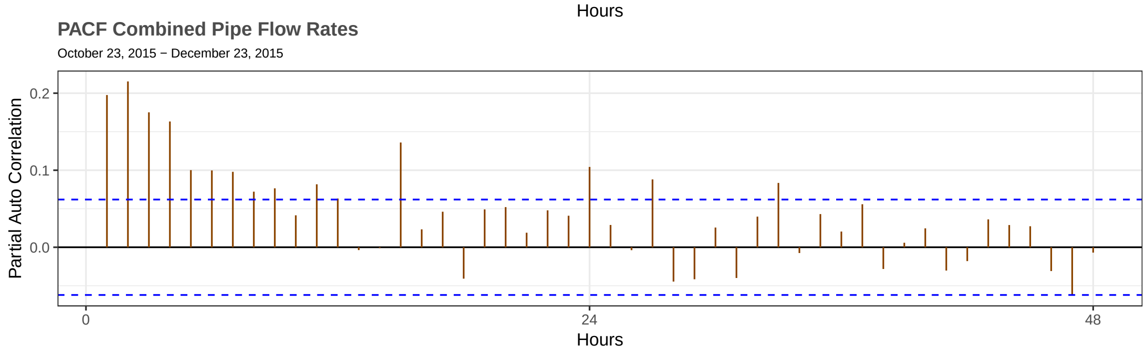
```
FALSE
FALSE   Augmented Dickey-Fuller Test
FALSE
FALSE data:  ws
FALSE Dickey-Fuller = -6.4409, Lag order = 9, p-value = 0.01
```

```
FALSE alternative hypothesis: stationary
```

Here we have contradictory esitmates, `ndiffs()` suggests a difference of 1, and the augmented dicky fuller test suggests that we are stationary as-is. An `auto.arima()` may give us a reasonable starting place.

1.3 Estimating Orders for ARIMA

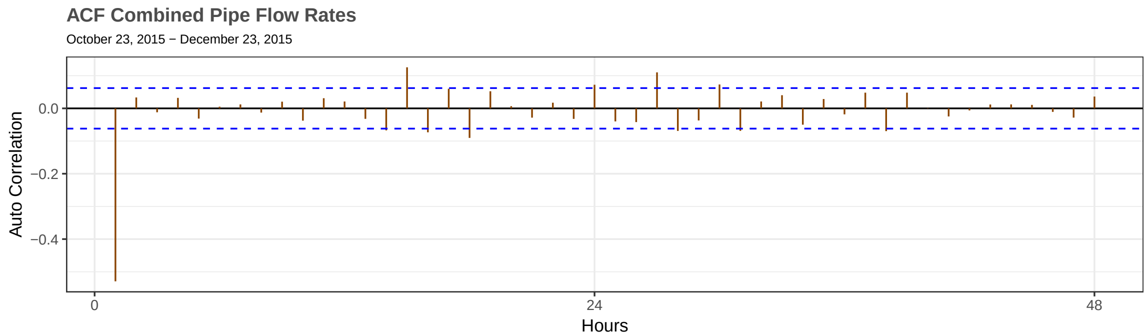


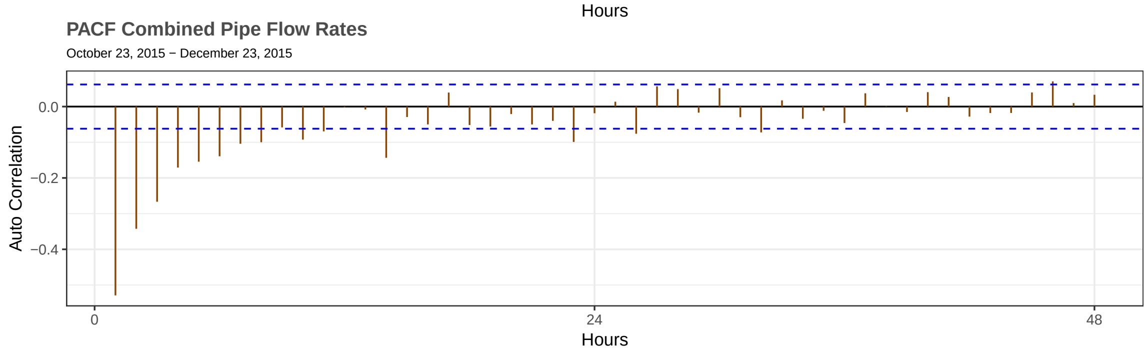


Interpreting the ACF and PACF

The ACF remain wholly above the critical threshold, so will likely require differencing as suggested by the `ndiffs()`, in looking at the PACF, there is some abiguity caused by the needed differencing, but after the intial trend down below the critical threshold, there is definitely a slight spike at 24, which would suggest there may indeed by a daily period or season we need to account for in our forecast.

Differenced ACF





A final ACF of the differenced data was done to ensure that a seconf first-order difference was not needed; thus we assume $d = 1$, a but it was not so clear about the appropriate value of $q4 should it be 5? , so `auto.arima()` is in order to help iterate up on the likely best starting place

1.4 `auto.arima()`

Using a Box-Cox lambda value to normalize the data may make $\lambda = .931552$. Because models can vary a lot based on the selection criterion, both BIC and AIC models were run, using lambda, to estimate a good starting place. We included the transformations in the model (instead of doing it outside the model), because we are using the ARIMA function to difference the data automatically allow more constiency and flexibility in testing other model orders.

The *AICc* chose a seasonal ARIMA of the following order:

$ARIMA(1, 1, 3)(0, 0, 1)[24]$ *AIC=7359.84 AICc=7359.9 BIC=7384.38*

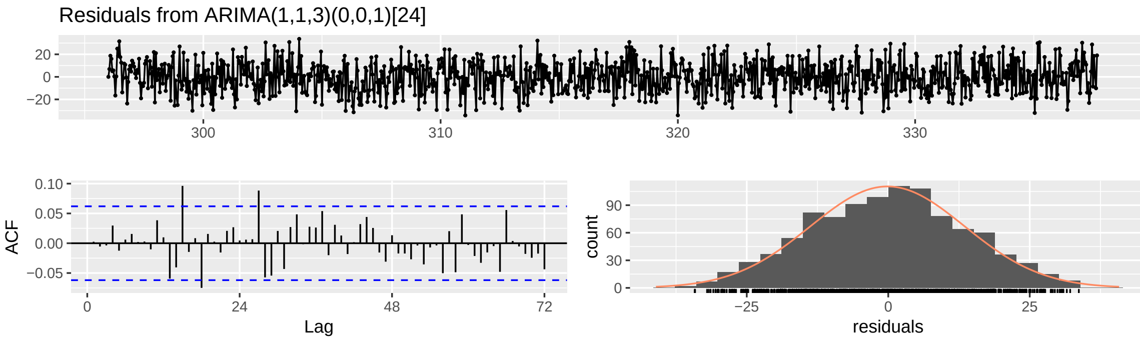
The *BIC* chose a non-seasonal ARIMA model as follows:

$ARIMA(2, 1, 1)$ *AIC=8082.22 AICc=8082.26 BIC=8101.85*

In both cases, the arima estimated that there needed to be differencing which was supported by `ndiffs()` and our ACF & PACF plots.

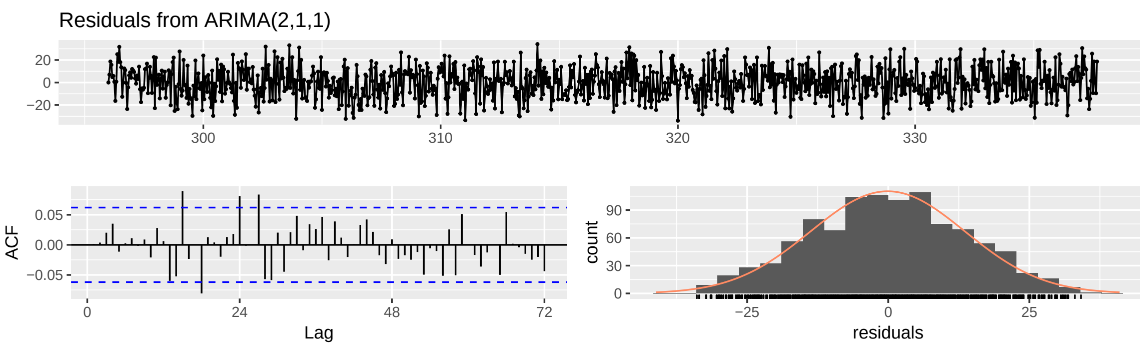
In comparing the two forecasts, for these automated models, they both degrade toward the series mean pretty quickly, however, the AICc model makes forecasts which consider the variation of the model a bit better before it levels out. So we decided to explore this model and see if we could tune it to provide more robust predictions

AIC $ARIMA(1,1,3)(0,0,1)[24]$ **Residual Plots**




```
FALSE
FALSE   Ljung-Box test
FALSE
FALSE data:  Residuals from ARIMA(1,1,3)(0,0,1)[24]
FALSE Q* = 57.362, df = 43, p-value = 0.07027
FALSE
FALSE Model df: 5.   Total lags used: 48
```

BIC $ARIMA(2,1,1)$ **Residual Plots**



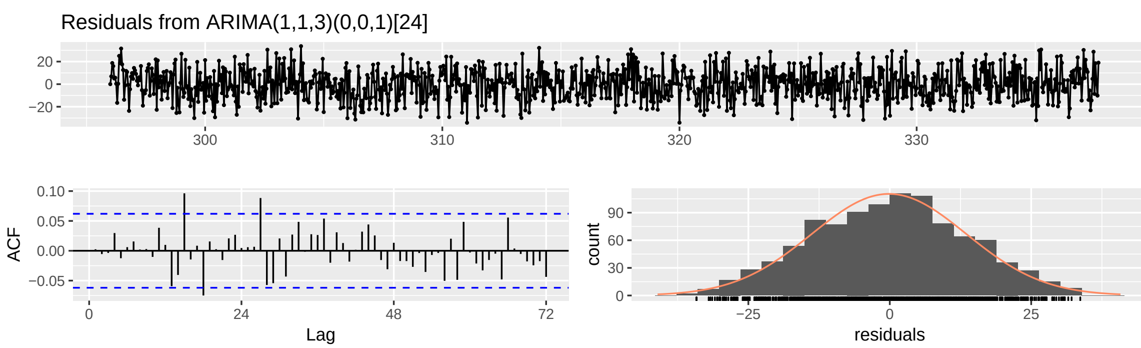

```
FALSE
FALSE   Ljung-Box test
FALSE
FALSE data:  Residuals from ARIMA(2,1,1)
FALSE Q* = 64.403, df = 45, p-value = 0.03029
FALSE
FALSE Model df: 3.   Total lags used: 48
```

1.4.1 Interpreting `auto.arima()`

In looking at the AICc and BIC ARIMA models, the both appear to be relatively white-noisy with no autocorrelation on the first or 24th observations, with relatively normal residuals. However, in looking at the Ljung-Box test for independence, it is clear that the Seasonal $ARIMA(1,1,3)(0,0,1)[24]$ is independent, where the $ARIMA(2,1,1)$ is not, thus reaffirming the lingering suspicion that thee is unaccounted for seasonal variation in the model requiring a seasona MA(1) to rectify. To be sure that the best model has been found, p & q as well as Q will be varied to see if a slight modification improves the performance of the model.

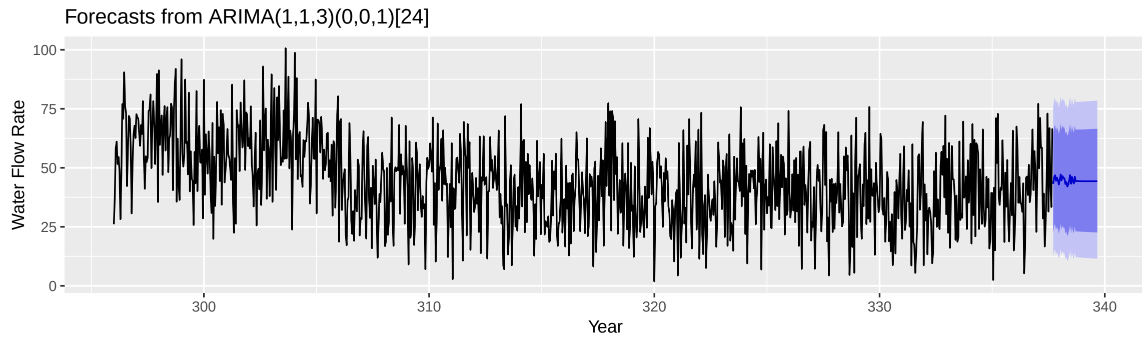
1.4.2 Manual ARIMA testing

```
FALSE Series: ws
FALSE ARIMA(1,1,3)(0,0,1)[24]
FALSE Box Cox transformation: lambda= 0.9531552
FALSE
FALSE Coefficients:
FALSE          ar1      ma1     ma2      ma3    sma1
FALSE       0.7602  -1.7578  0.8286  -0.0614  0.0833
FALSE s.e.  0.1857   0.1874  0.1886   0.0324  0.0320
FALSE
FALSE sigma^2 estimated as 187:  log likelihood=-4033.28
FALSE AIC=8078.56   AICc=8078.64   BIC=8108
```



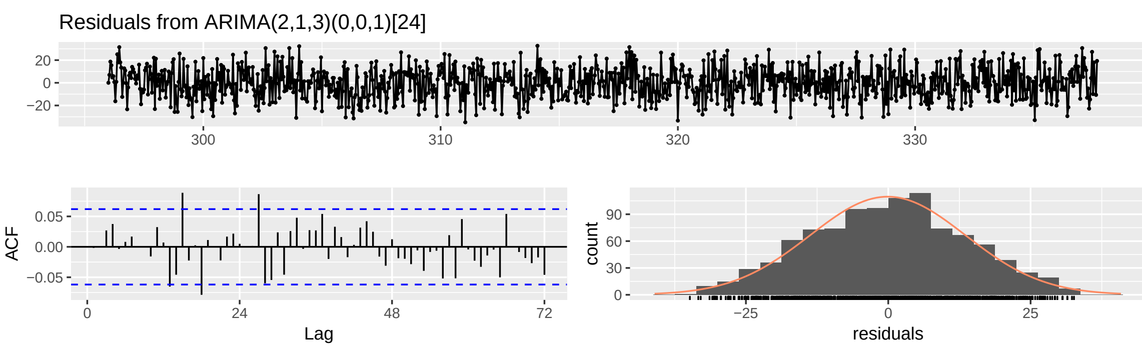

```
FALSE
FALSE   Ljung-Box test
FALSE
FALSE data:  Residuals from ARIMA(1,1,3)(0,0,1)[24]
FALSE Q* = 47.142, df = 31, p-value = 0.03174
FALSE
FALSE Model df: 5.   Total lags used: 36
```

1.4.2.1 Forecasting From the ARIMA



1.4.2.2 $ARIMA(2,1,3)(0,0,1)[24]$

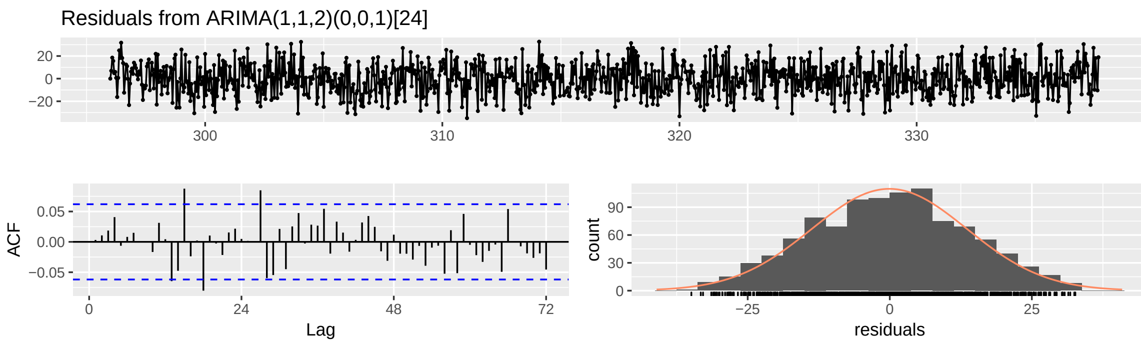
```
FALSE Series: ws
FALSE ARIMA(2,1,3)(0,0,1)[24]
FALSE Box Cox transformation: lambda= 0.9531552
FALSE
FALSE Coefficients:
FALSE           ar1     ar2      ma1      ma2     ma3    sma1
FALSE       -0.1435  0.1884  -0.8478  -0.2709  0.1621  0.0798
FALSE s.e.      NaN  0.5408      NaN   0.6070  0.5319  0.0318
FALSE
FALSE sigma^2 estimated as 187.5:  log likelihood=-4034.02
FALSE AIC=8082.05   AICc=8082.16   BIC=8116.4
```




```
FALSE
FALSE   Ljung-Box test
FALSE
FALSE data:  Residuals from ARIMA(2,1,3)(0,0,1)[24]
FALSE Q* = 48.506, df = 30, p-value = 0.01764
FALSE
FALSE Model df: 6.   Total lags used: 36
```

This Ljung-Box shows unexplained variances in the residuals indicating that this model is not yet fully realized and inferior to the Seasonal $ARIMA(1,1,3)(0,0,1)[24]$.

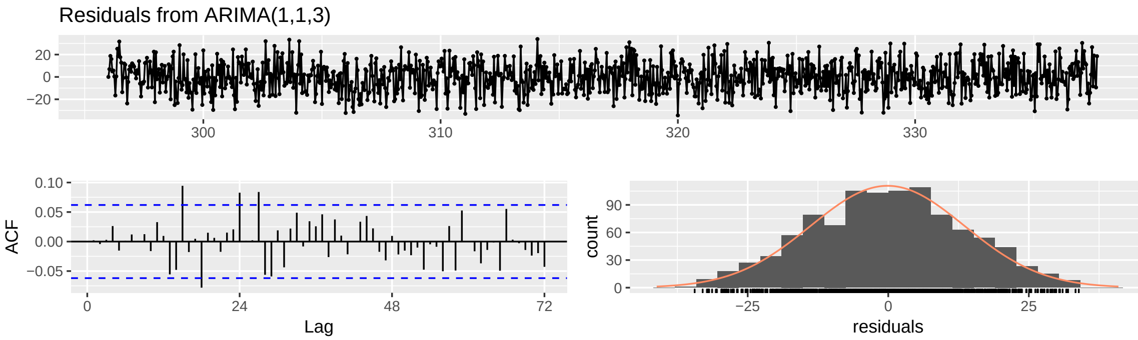
```
FALSE Series: ws
FALSE ARIMA(1,1,2)(0,0,1)[24]
FALSE Box Cox transformation: lambda= 0.9531552
FALSE
FALSE Coefficients:
FALSE           ar1      ma1      ma2    sma1
FALSE       -0.2655  -0.7307  -0.2103  0.0790
FALSE s.e.   0.9490   0.9533   0.9121  0.0318
FALSE
FALSE sigma^2 estimated as 187.1:  log likelihood=-4034.08
FALSE AIC=8078.16   AICc=8078.22   BIC=8102.7
```




```
FALSE
FALSE   Ljung-Box test
FALSE
FALSE data:  Residuals from ARIMA(1,1,2)(0,0,1)[24]
FALSE Q* = 47.963, df = 32, p-value = 0.03467
FALSE
FALSE Model df: 4.   Total lags used: 36
```

This Ljung-Box also shows unexplained variances in the residuals indicating that this model is not yet fully realized and inferior to the Seasonal $ARIMA(1,1,2)(0,0,1)[24]$.

```
FALSE Series: ws
FALSE ARIMA(1,1,3)
FALSE Box Cox transformation: lambda= 0.9531552
FALSE
FALSE Coefficients:
FALSE          ar1      ma1     ma2      ma3
FALSE       0.6792  -1.6742  0.7437  -0.0553
FALSE s.e.  0.2923   0.2930  0.2903   0.0330
FALSE
FALSE sigma^2 estimated as 188.1:  log likelihood=-4036.63
FALSE AIC=8083.27   AICc=8083.33   BIC=8107.81
```

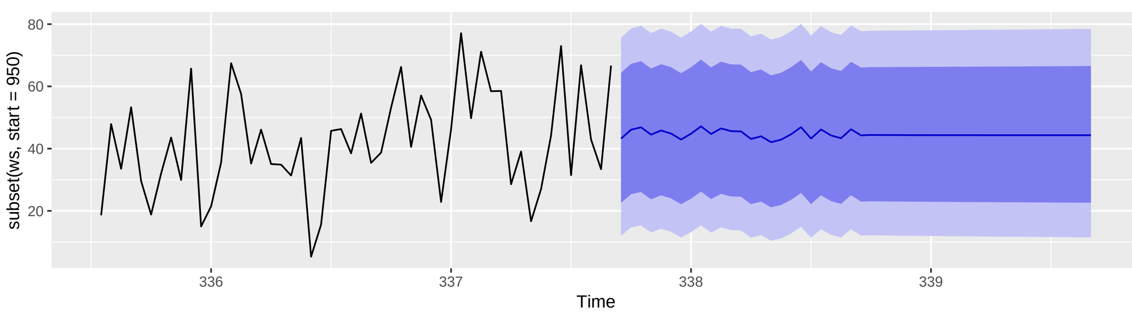



```
FALSE
FALSE   Ljung-Box test
FALSE
FALSE data:  Residuals from ARIMA(1,1,3)
FALSE Q* = 53.61, df = 32, p-value = 0.009708
FALSE
FALSE Model df: 4.   Total lags used: 36
```

This Ljung-Box also shows unexplained variances in the residuals indicating that this model is not yet fully realized and inferior to the Seasonal $ARIMA(1,1,3)$.

1.4.3 Accepting the `auto.arima()`

Given that the other models show unexplained variance in the residuals, the final predictions will be made using the AICc recommended model of $ARIMA(1,1,3)(0,0,1)[24]$.



Sample Forecasts

1.4.3.1 Forecast Accuracy

	ME	RMSE	MAE	MPE	MAPE	MASE	ACF1
Training set	0.0015679	16.27402	13.23093	-28.76247	50.34448	0.7489308	0.0014339

Table 1.1: First few predictions in the set

DateTime	Point Forecast	Lo 80	Hi 80	Lo 95	Hi 95
2015-12-03 17:00:00	43.21837	22.59441	64.33311	12.00034	75.65243
2015-12-03 18:00:00	46.07958	25.37341	67.24682	14.70394	78.58795
2015-12-03 19:00:00	46.85016	26.06919	68.08732	15.35468	79.46457
2015-12-03 20:00:00	44.49638	23.73897	65.73546	13.06315	77.11903
2015-12-03 21:00:00	45.83029	25.00018	67.13008	14.27275	78.54342
2015-12-03 22:00:00	44.85032	24.01864	66.16308	13.30217	77.58566

1.5 Summary

Ultimately this model is marginally useful as seen by the Mean Absolute Percentage of Error which reveals that the average percentage each forecast is off by is around 50%. In looking at the graph of the forecast above, which is the last 150 points in the time series and the forecasted points, you can see this as the predictions lightly modulate around the mean and deteriorate to it pretty quickly.

In looking at the original decomposition, there very little trend, a lot of seasonality, is a pretty substatial amount of random noise, which is not considered in the model, and is responsible for the majority of the error in this model, as white noise is never predictable.