Final Project Draft

James King & Joe Dickinson

3/15/2022

Bend, Oregon

Introduction

One of the group members, James, enjoys trail running in Deschutes County, OR, so was inspired to put together a dataset to explore the Air Quality Index (AQI) throughout the seasons and years since summer wildfire season is becoming more extreme in the Pacific Northwest. The dataset used for this project was put together by combining annual "Daily AQI by County" datasets which are published by the EPA.

The dataset for this project is daily high observations of the Air Quality Index for Deschutes County, Oregon from January 2012 through June 2021 (3,428 total observations). The original dataset contains many attributes - including attributes for state, county, state and county codes, etc - but for this project only AQI and date are considered.

Methods

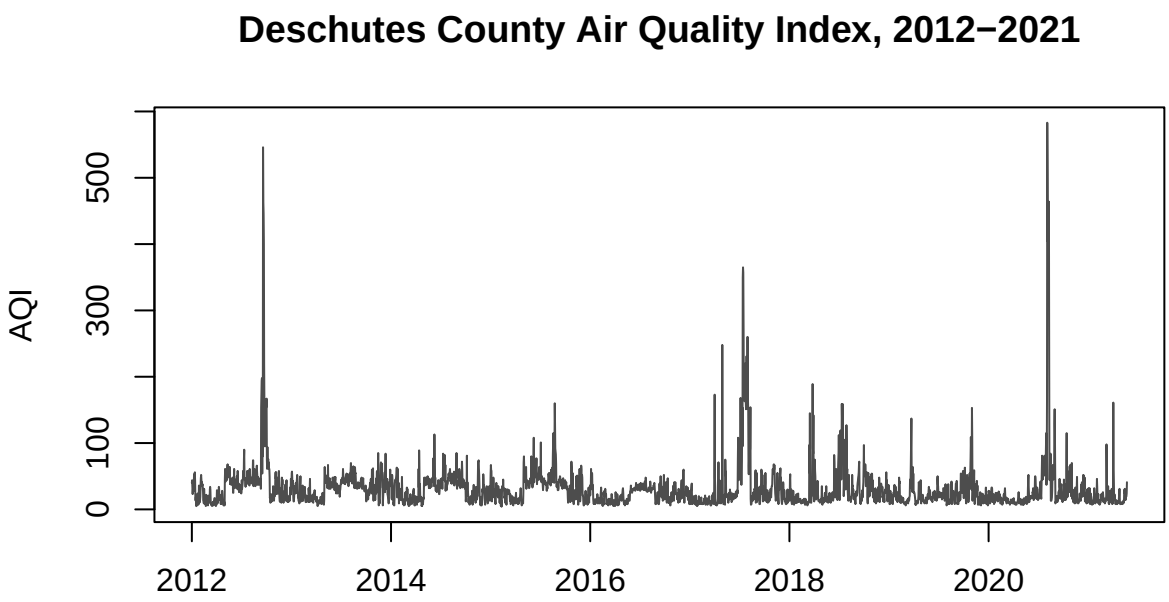

Figure 1

Looking at a plot of the data (Figure 1), the AQI values range from 4 to 583. A higher AQI value indicates lower air quality. The bulk of observations are less than 100 and the dataset has a mean of 31.78 and a median of 23. There are a handful of extreme values exceeding 200. All of this indicates skewness in the data.

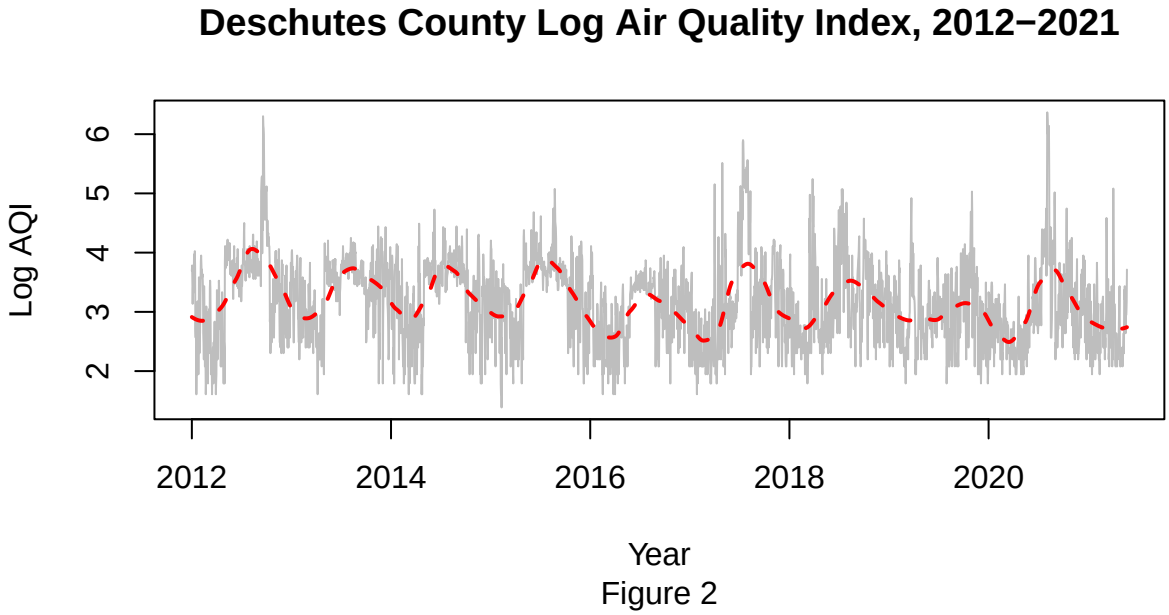

Figure 2

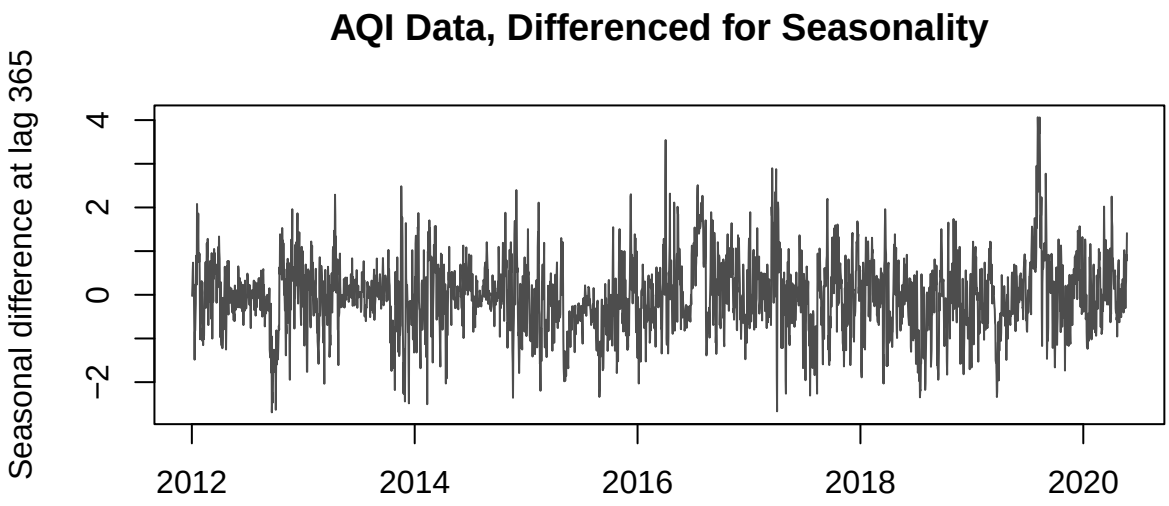

Figure 3

A log transformation is applied to the data to offset the skewness. Looking at the plot of the transformed data (Figure 2) and applying a smoothed trend line, it is clear there is annual seasonality to the data. This is expected as large forest fires tend to happen in the late summer and early fall. There appears to be no strong trend in the data, although a case could be made for increasing variance.

Running a Dickey-Fuller test on the data and the log-transformed data show both are stationary.

Figure 3 shows the data after differencing is applied to remove the seasonality. A lag of 365 was used to indicate the yearly seasonality of the daily observations.

To arrive at a standard arima model as covered within the scope of this course, we examined auto-correlation function (ACF) and partial auto-correlation function (PACF) plots on the log transformed original time series and on the log transformed time series with seasonal differencing applied. The plots (Figure 4) suggest a better fit using the log transformed time series with seasonal differencing.

Once our model was fit, we examined the ACF and PACF plots of the residuals, used the diagnostic tools in the `tsdiag` function, and examined a QQ plot of the residuals in order to assess the appropriateness of the model fit. Finally, we forecasted values and back-transformed the results onto the original time series using the `predict` function. This is shown in Figures 5-8.

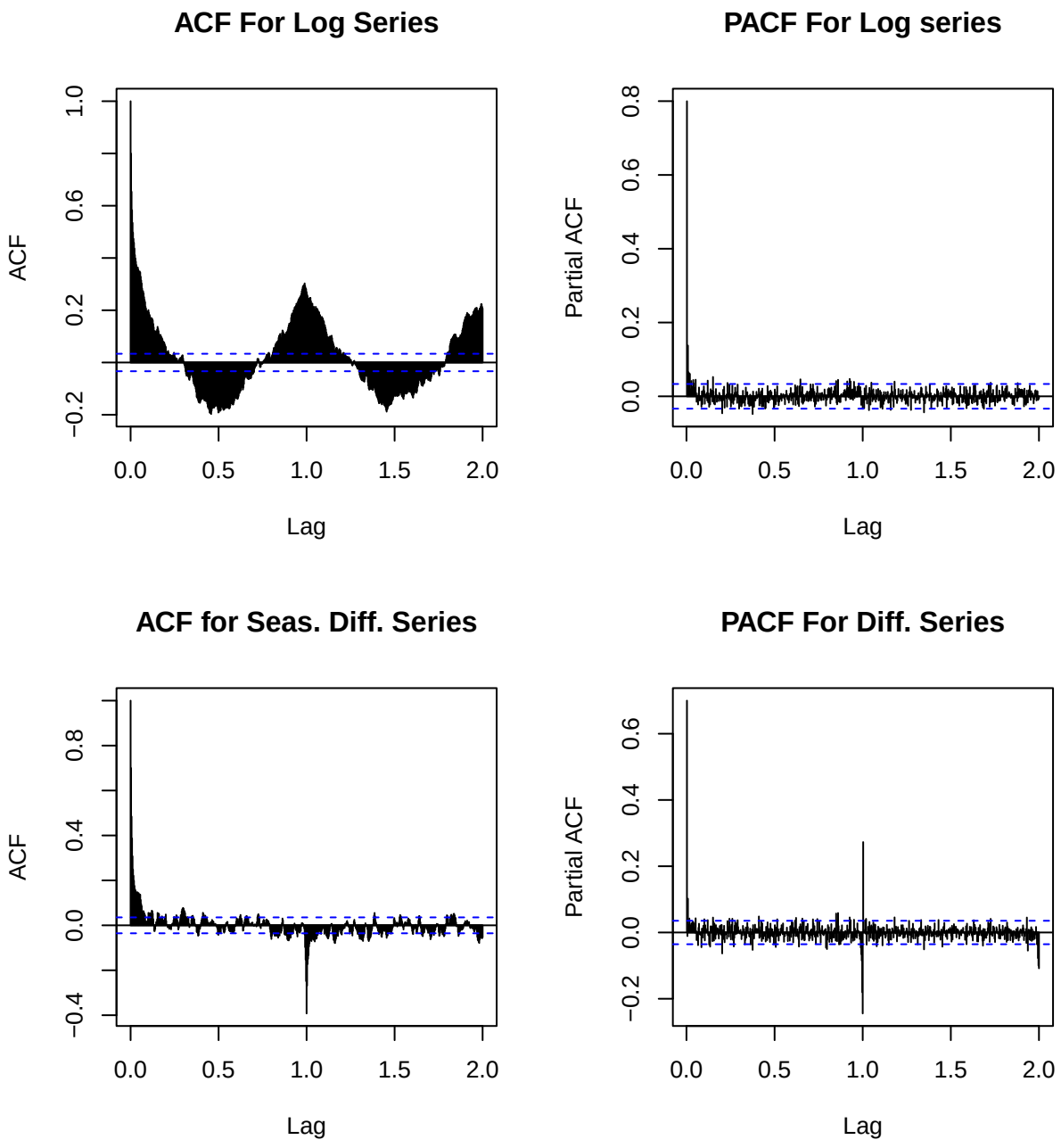
ACF For Log Series
ACF
1.0
0.6
0.2
−0.2
0.0
0.5
1.0
1.5
2.0
Lag
PACF For Log series
Partial ACF
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
Lag
ACF for Seas. Diff. Series
ACF
0.8
0.4
0.0
−0.4
0.0
0.5
1.0
1.5
2.0
Lag
PACF For Diff. Series
Partial ACF
0.6
0.4
0.2
0.0
−0.2
0.0
0.5
1.0
1.5
2.0
Lag

Figure 4

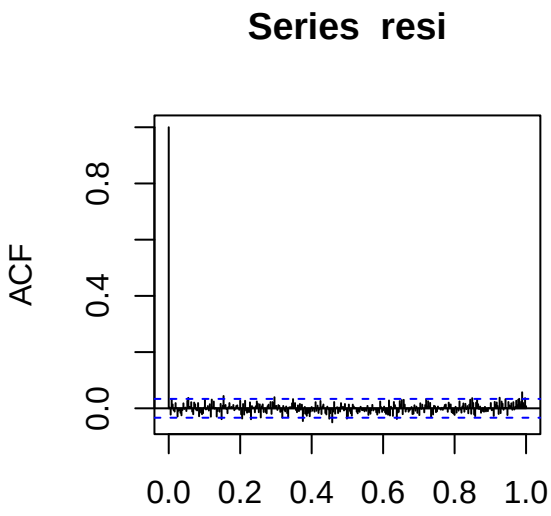

Figure 5

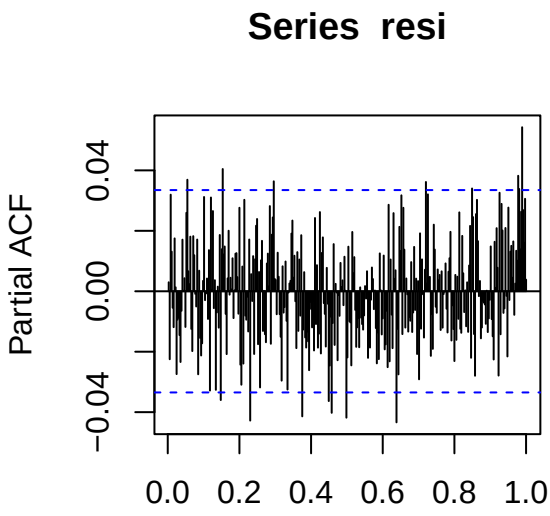

Figure 6

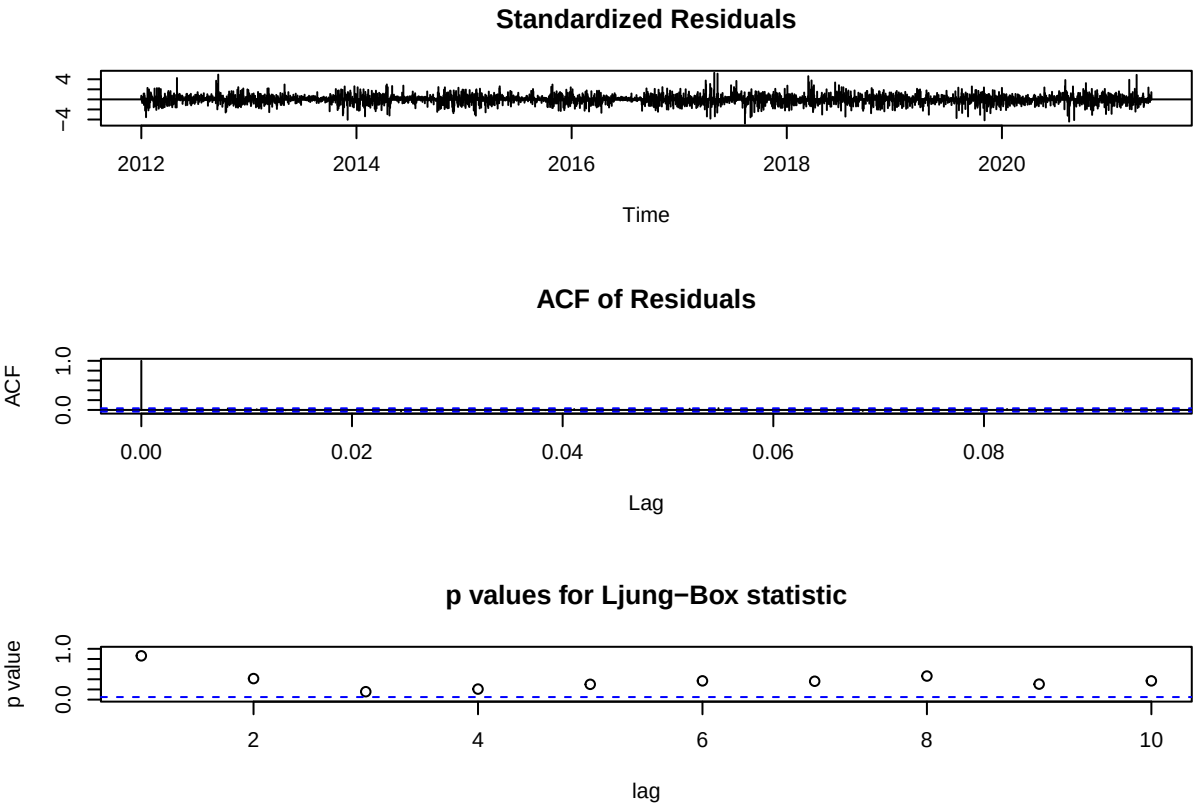

Figure 7

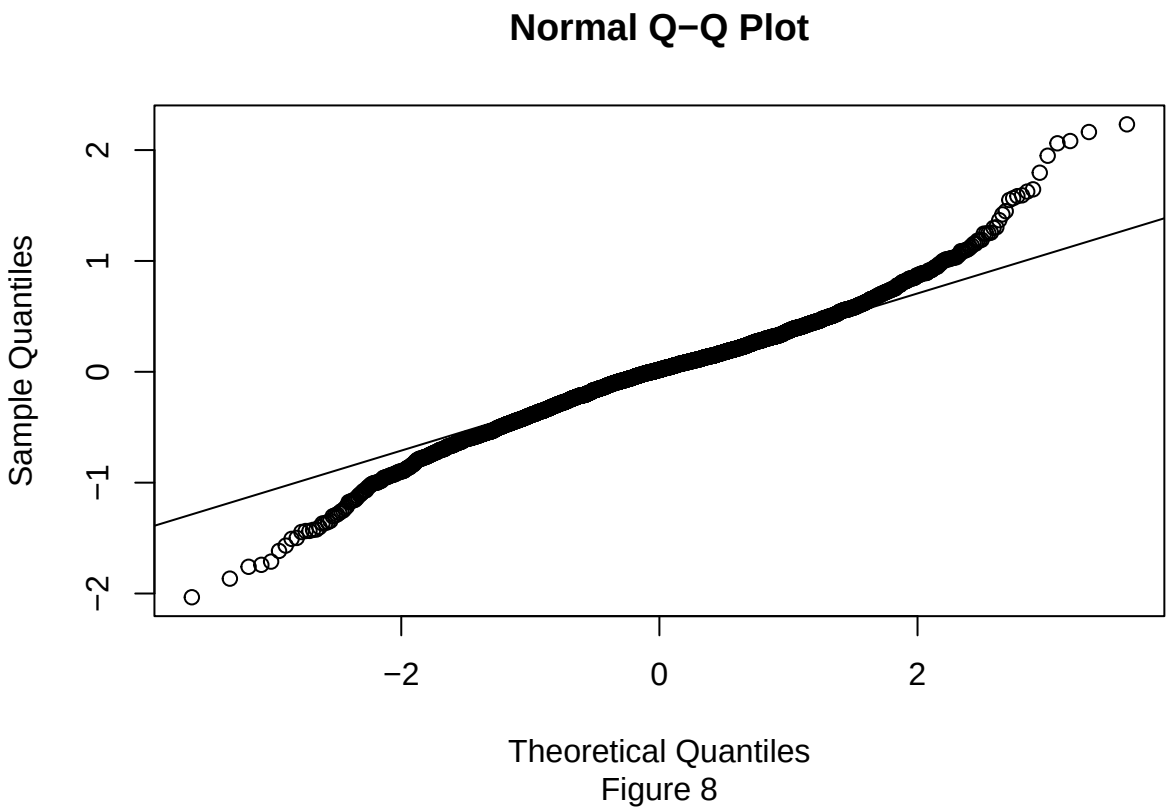

Figure 8

Given the seasonal limitations of ARIMA models in R, we sought a more sophisticated modeling approach. The TBATS model from the `forecast` package incorporates multiple tools for modeling time series data. It uses Box-Cox transformation, ARIMA, trigonometric seasonality, and exponential smoothing. TBATS is most useful for data with multiple seasonality and while this dataset likely has only annual seasonality, the combination of methods is well suited for the data.

Data	AIC
Original AQI	43705.68
Log AQI	22020.64
Differenced Log AQI	21358.83

Results

In examining the log transformed time series, it is apparent seasonality is present. When the time series is differenced, seasonality is still present. When a seasonal difference (lag = 365) is applied to the log transformed time series, much of the seasonality is removed.

In running the `auto.arima` function on each time series, we get a lower AIC value on the log transformed time series (AIC = 3889 on an ARMA(3, 1, 1)) model as opposed to the model built on the seasonally differenced time series (AIC = 5480 on an ARMA(3, 0, 0) model). This was unexpected as there is seasonality present in the data, however we proceeded with the log transformed data without seasonal differencing.

We believe further optimization could be done with knowledge outside the scope of the course. This is why the TBATS method was explored. In trying to build a SARIMA model, or using the `seasonal` call within the `arima` function, we ran into issues where the function would not support using a lag > 350. If the function allowed a lag of 365, we would have tried a (p, d, q) x (P, D, Q) model of (3, 1, 1) x (1, 1, 1), however.

The final ARMA(3, 1, 1) model on the log transformed time series had an AIC value of 3837. The ACF and PACF plots of the residuals suggested some seasonality present. In the `tsdiag` diagnostic plots we see that there were a few outlier data points across the time series, however the p-values for the Ljung-Box statistic plot were favorable with all being > 0.05. The QQ plot indicates normality, with the exception of some outliers at the extremes. The model fit appears to be appropriate for our purposes.

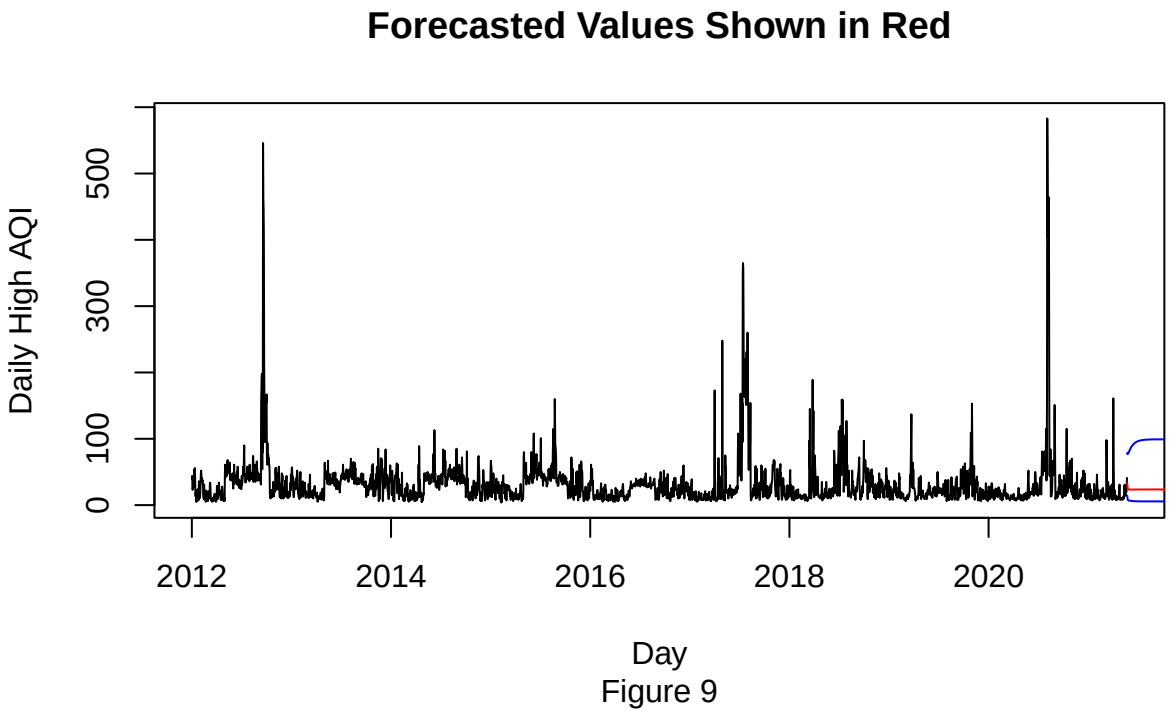

Figure 9

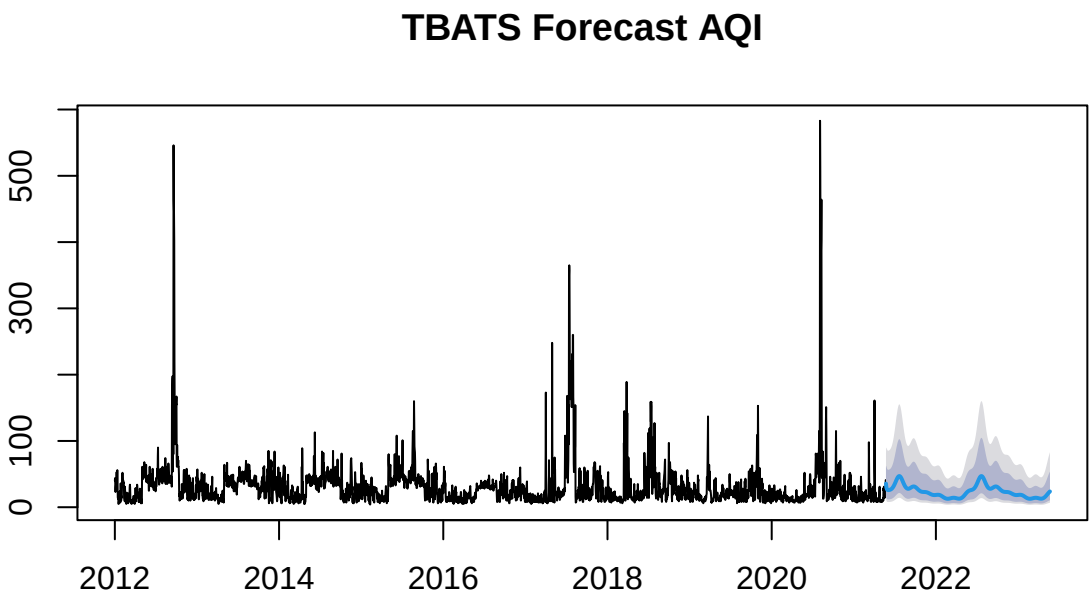

Figure 10

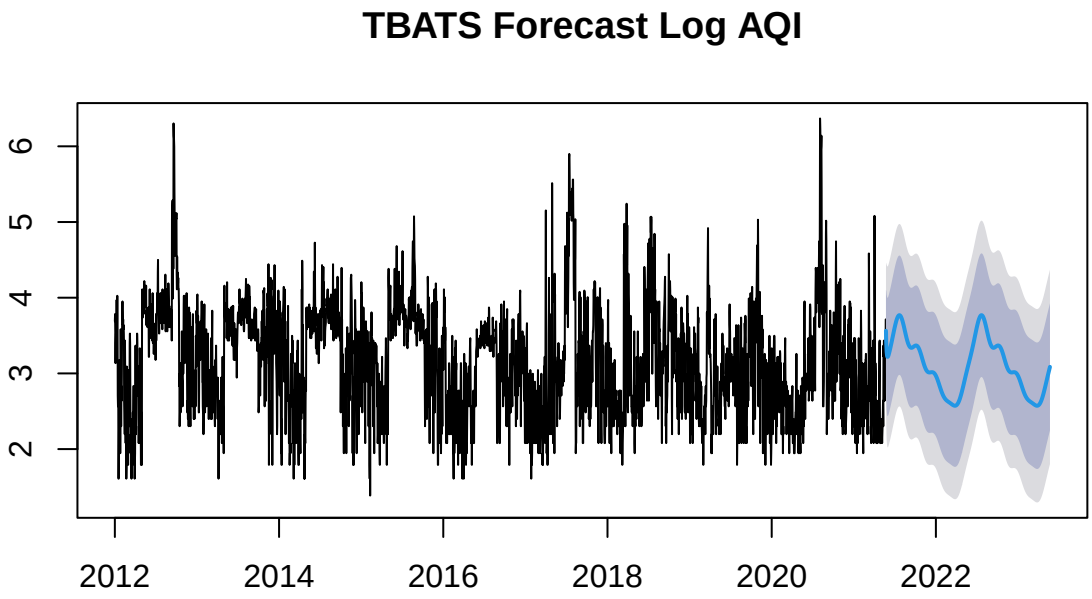

Figure 11

Using the `predict` function on this model, we get a linear average forecast which aligns with the average of the entire time series, with wide confidence intervals (Figure 9). The TBATS forecasts for AQI and log AQI show a continuing pattern of higher AQI in the summer months trailing off toward winter (Figures 10 and 11).

Discussion

We believe this dataset provides a good example of where more sophisticated forecasting procedures may be useful. Using the methods learned in this course, the forecasted values do not ebb and flow with the seasons. However, the TBATS forecasting method does, and provides reasonable and intuitive-looking results.

Some observations to note about the time series is that the optimized model identified by the TBATS tool did not identify seasonality as being present. When taking a closer look at the original time series, it does appear that seasonality is less apparent in more recent years. Air quality does appear to be getting slightly better on average over time, but with more extreme isolated events. A further look at these data with a better understanding of more sophisticated forecasting procedures could prove to be more insightful.

Reference

https://aqs.epa.gov/aqsweb/airdata/download_files.html#AQI

Appendix

```
knitr::opts_chunk$set(echo = FALSE, warning=FALSE, cache = TRUE)
library(forecast)
library(tseries)
knitr::include_graphics("800px-BendORPanoramic.jpg")
#read in data, turn into time series, run median and mean
aqi_data <- read.csv('daily_aqi_deschutes_county_2012-2021.csv',header = TRUE)
aqi_ts <- ts(aqi_data$AQI,start = c(2012,1,1), frequency = 365)
median(aqi_ts)
mean(aqi_ts)
plot(aqi_ts, col = "gray30", xlab = "Year", ylab = "AQI",
     main = "Deschutes County Air Quality Index, 2012-2021", sub = "Figure 1")
#log transform data, plot with smoothed trend line
aqi_log <- log(aqi_ts)
aqi_time <- time(aqi_log)
aqi_loess <- loess(aqi_log ~ aqi_time, span = 0.1)
aqi_loess_pred <- predict(aqi_loess)
aqi_loess_trend <- ts(aqi_loess_pred,start = c(2012,1,1), frequency = 365)
#Dickey-Fuller test
adf.test(aqi_ts)
adf.test(aqi_log)
#difference for seasonality
diff_yr <- c(NA, diff(aqi_log, lag = 365))
diff_yr <- ts(diff_yr,start = c(2012,1,1), frequency = 365)
#plots for log tranformed and seasonally differenced data
##par(mfrow = c(2,1))
plot(aqi_log, col = "gray", xlab = "Year", ylab = "Log AQI",
     main = "Deschutes County Log Air Quality Index, 2012-2021", sub = "Figure 2")
lines(aqi_loess_trend,col="red",lty=2,lwd=2)
plot(diff_yr, col = "gray30", ylab = "Seasonal difference at lag 365",
     main = "AQI Data, Differenced for Seasonality", sub = "Figure 3")

par(mfrow = c(2, 2))

# ACF and PACF plots on log data
acf(aqi_log, lag.max = 730, na.action = na.pass,
    main = "ACF For Log Series")
pacf(aqi_log, lag.max = 730, na.action = na.pass,
     main = "PACF For Log series")

# ACF and PACF plots on seasonal diff of log data
acf(diff_yr, lag.max = 730, na.action = na.pass,
    main = "ACF for Seas. Diff. Series")
pacf(diff_yr, lag.max = 730, na.action = na.pass,
     main = "PACF For Diff. Series")
# Use `auto.arima` function to verify model selection
log.auto <- auto.arima(aqi_log)
sdiff.auto <- auto.arima(diff_yr)

log.auto
sdiff.auto
# Keep this as an example of what we tried but didn't work?
```

```r
# Doesn't work
# test.fit <- arima(aqi.diff1,
#                   order = c(3, 1, 1),
#                   seasonal = list(order = c(1, 1, 1),
#                                   period = 365))
# test.fit

# Doesn't work
# test.fit2 <- arima(aqi.log,
#                    order = c(3, 1, 1),
#                    seasonal = list(order = c(1, 1, 1),
#                                    period = 365))
# test.fit2

# Doesn't work
# library(astsa)
# test.fit3 <- sarima(aqi.log,
#                     3, 0, 1,
#                     P = 1, D = 0, Q = 1,
#                     S = 365)
# test.fit3
# Fit the model with the lowest AIC value
fit.arma <- arima(aqi_log,
                  order = c(3, 0, 1),
                  seasonal = list(order = c(0, 0, 0),
                                  period = 365))
fit.arma
# Fit model on seasonal diff of log data
fit.arma.s <- arima(diff_yr,
                  order = c(3, 0, 0),
                  seasonal = list(order = c(0, 0, 0),
                                  period = 365))
fit.arma.s
##### Figures 5 and 6 #####
par(mfrow = c(1, 2))

# Fit the residuals
resi <- fit.arma$residuals
resi <- na.omit(resi)

acf(resi, lag.max = 365, sub = "Figure 5")
pacf(resi, lag.max = 365, sub = "Figure 6")
##### Figure 7 ##### I can't figure out how to add a subtitle here
# Diagnostic plots
tsdiag(fit.arma)
##### Figure 8 #####
# QQ plot
qqnorm(resi, sub = "Figure 8")
qqline(resi)
# Ru
aqi_tbats <- tbats(aqi_ts)
aqi_log_tbats <- tbats(aqi_log)
aqi_diff_yr_tbats <- tbats(diff_yr)
```

```r
aqi_aic <- format(round(aqi_tbats$AIC,2), scientific = FALSE)
aqi_log_aic <- format(round(aqi_log_tbats$AIC,2), scientific = FALSE)
aqi_diff_aic <- format(round(aqi_diff_yr_tbats$AIC,2), scientific = FALSE)
# Forecast log transformed data
pred <- predict(fit.arma, n.ahead = 365)

plot(aqi_log,
     main = "Forecasted Values Shown in Red",
     xlab = "Day",
     ylab = "Log(Daily High AQI)")

# Forecasted values
lines(pred$pred, col = "red")

# 95% forecasting limits
lines(pred$pred - 2 * pred$se, col = 'blue')
lines(pred$pred + 2 * pred$se, col = 'blue')
##### Figure 9 #####
# Forecast for original data
plot(aqi_ts,
     main = "Forecasted Values Shown in Red",
     xlab = "Day",
     ylab = "Daily High AQI",
     sub = "Figure 9")

# forecasted values
lines(exp(pred$pred), col = "red")

# 95% forecasting limits
lines(exp(pred$pred - 2 * pred$se), col = 'blue')
lines(exp(pred$pred + 2 * pred$se), col = 'blue')
plot(forecast(aqi_tbats), main="TBATS Forecast AQI", sub = "Figure 10")
plot(forecast(aqi_log_tbats), main="TBATS Forecast Log AQI", sub = "Figure 11")
```