Time Series Project-2

Abhiruchi Shukla

12/20/2018

# Seasonal Time Series Analysis of Lake Erie water level

## Contents

# 1. Introduction

The Great Lakes are a series of interconnected fresh water lakes in Northern America. The Great Lakes are the largest group of freshwater lakes on Earth by total area, and second largest by total volume, containing 21% of the world's surface fresh water by volume. They consist of Lakes Superior, Michigan, Huron, Erie and Ontario. These lakes have been a major source of habitat and biodiversity in the surrounding region.

The surface water levels of these lakes fluctuate in response to a variety of factors. These fluctuations can have a positive or a negative impact on water dependent industries. Therefore, it is important to build a prediction model which can be used by these industries for planning and management purposes. The Great Lakes Environmental Research Laboratory has built extensive models to gain understanding of the water levels in a variety of ways.

In this project, I have performed a time series statistical analysis for the design of a model that could predict future water level based on the previous water levels using R. The data was collected on a monthly basis from January 1921 to December 1970 for Lake Erie water levels in tens of meters.

Dataset source: https://datamarket.com/data/set/22pw/monthly-lake-erie-levels-1921-1970#!ds=22pw&display=line

# 2. Model Building

## 2.1 Fit Data as Time Series

The time series, ACF and PACF plot is shown in Figure 1. The data appears to be non-. stationary with seasonal peaks. The ACF plots shows a sinusoidal pattern which decays slowly.

The ADF test yields a test statistic -2.3493 with p-value 0.4305, which implies that the model is non-stationary. Therefore, the data needs regular differencing. The Box and Cox transformation gave λ = -0.0103. Therefore, I performed a log transformation of the original time series data in order to achieve variance stability.

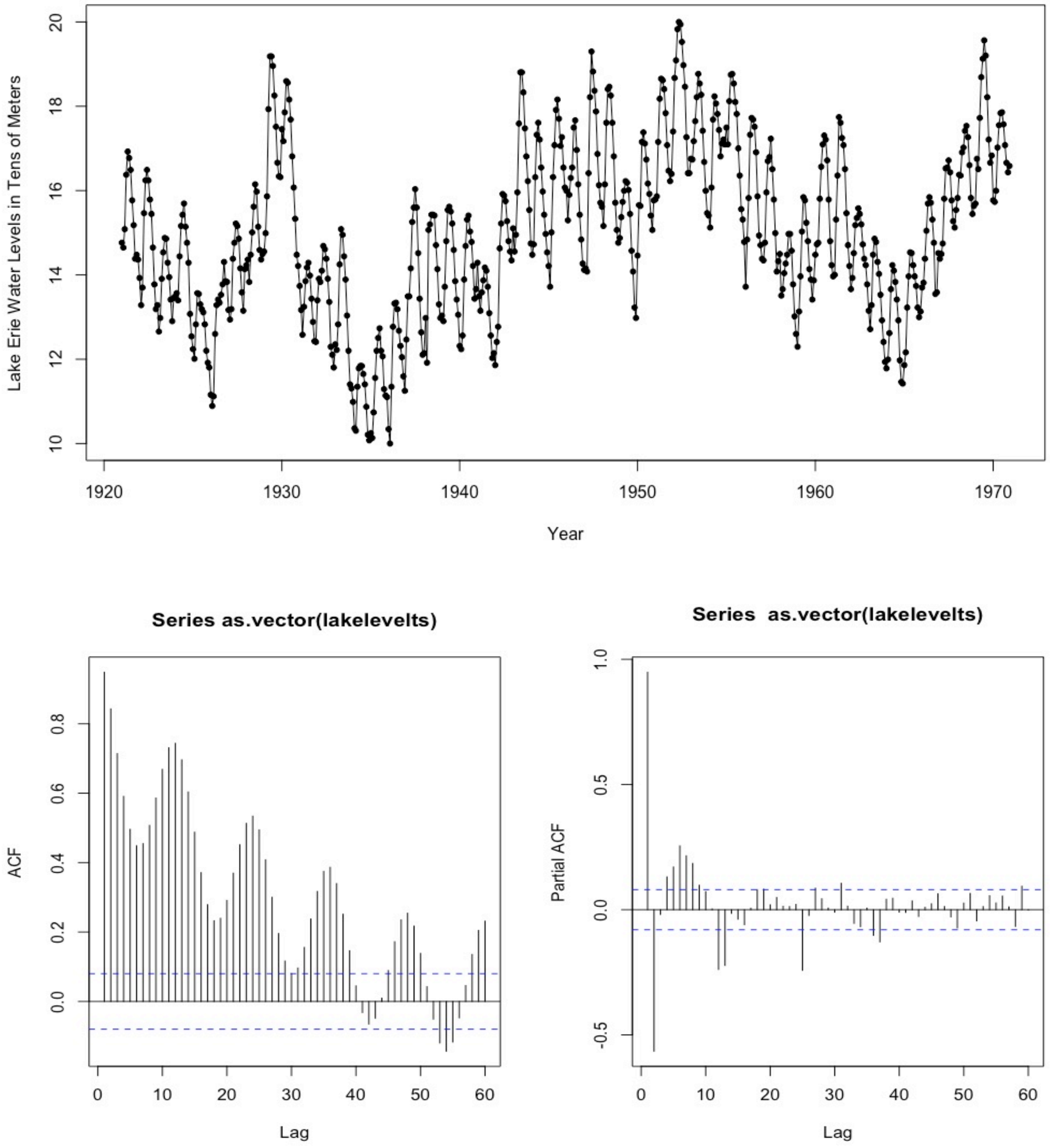

Figure 1: Time Series, ACF and PACF plot of the original time series data

## 2.2 Log transformed and differenced time series analysis

### 2.2.1 Q-q plot, ACF, and PACF of log transformed data

After taking the log transformation, I obtained a Q-Q plot of the transformed time series. The Q-Q plot, ACF and PACF of the transformed data are shown in Figure 2. The Q-Q plot corroborates assumption of normal distribution of data. From the ACF plots, it appears there are seasonal patterns present in the data, but we need to analyze further in order to confirm seasonality.

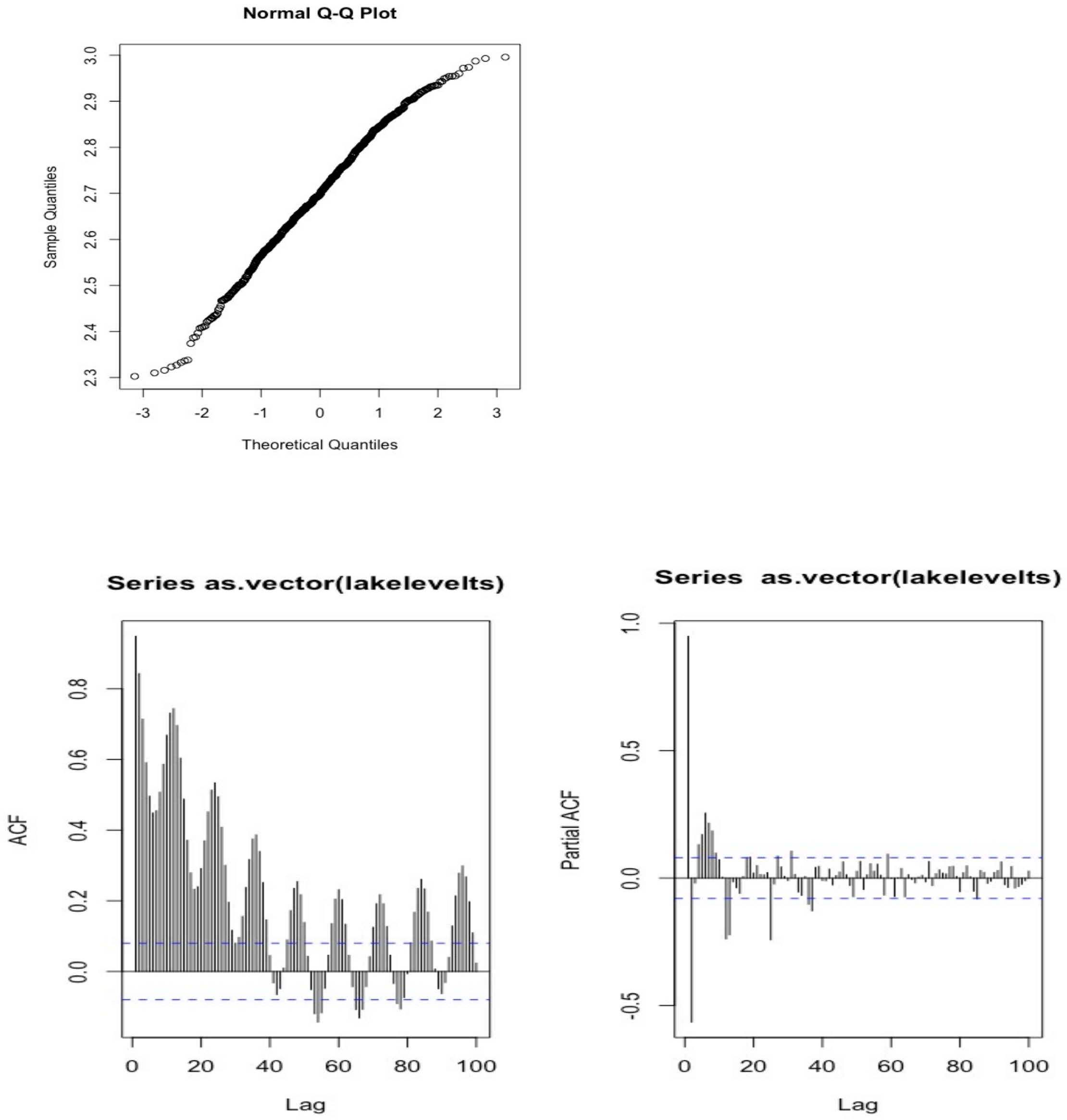

Figure 2: Q-Q plot, ACF and PACF for log transformed time series

### 2.2.2 Additive Decomposition to detect trend, seasonality, and randomness

Figure 3 shows the decomposition of the log transformed time series for trend, seasonal, and random components. It can be inferred that the transformed data needs to be seasonally differenced to account for seasonality and regularly differenced to account for trend. A box-plot was also obtained to study deviation of index values from centered averages. Clearly, there is a seasonality factor in the months from April-June.

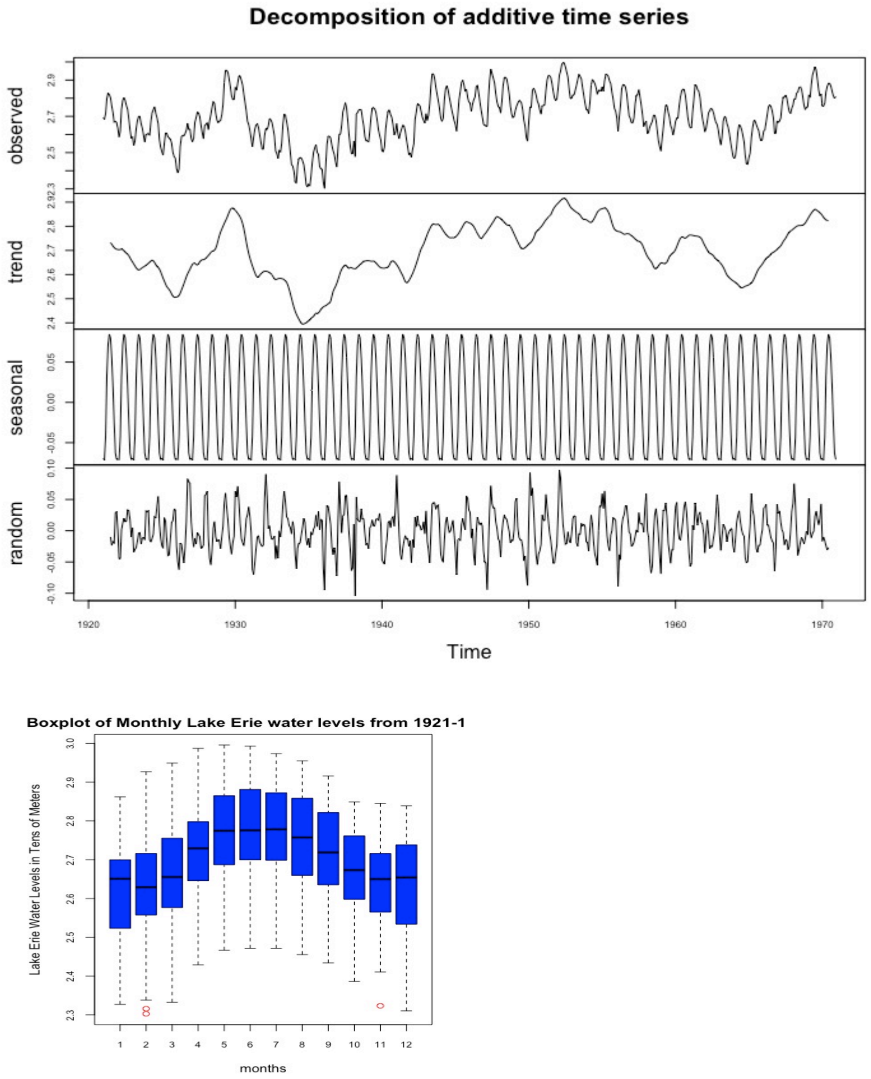

Figure 3: Additive decomposition of time series, and Box-plot

### 2.2.3 Periodogram

A periodogram was used to resolve the dominant frequencies present in the time-series to test particularly when the cycles are not related to the commonly encountered monthly or quarterly seasonality. Figure 4 shows the resulting periodogram obtained after regular differencing of transformed data along with ACF and PACF plots.

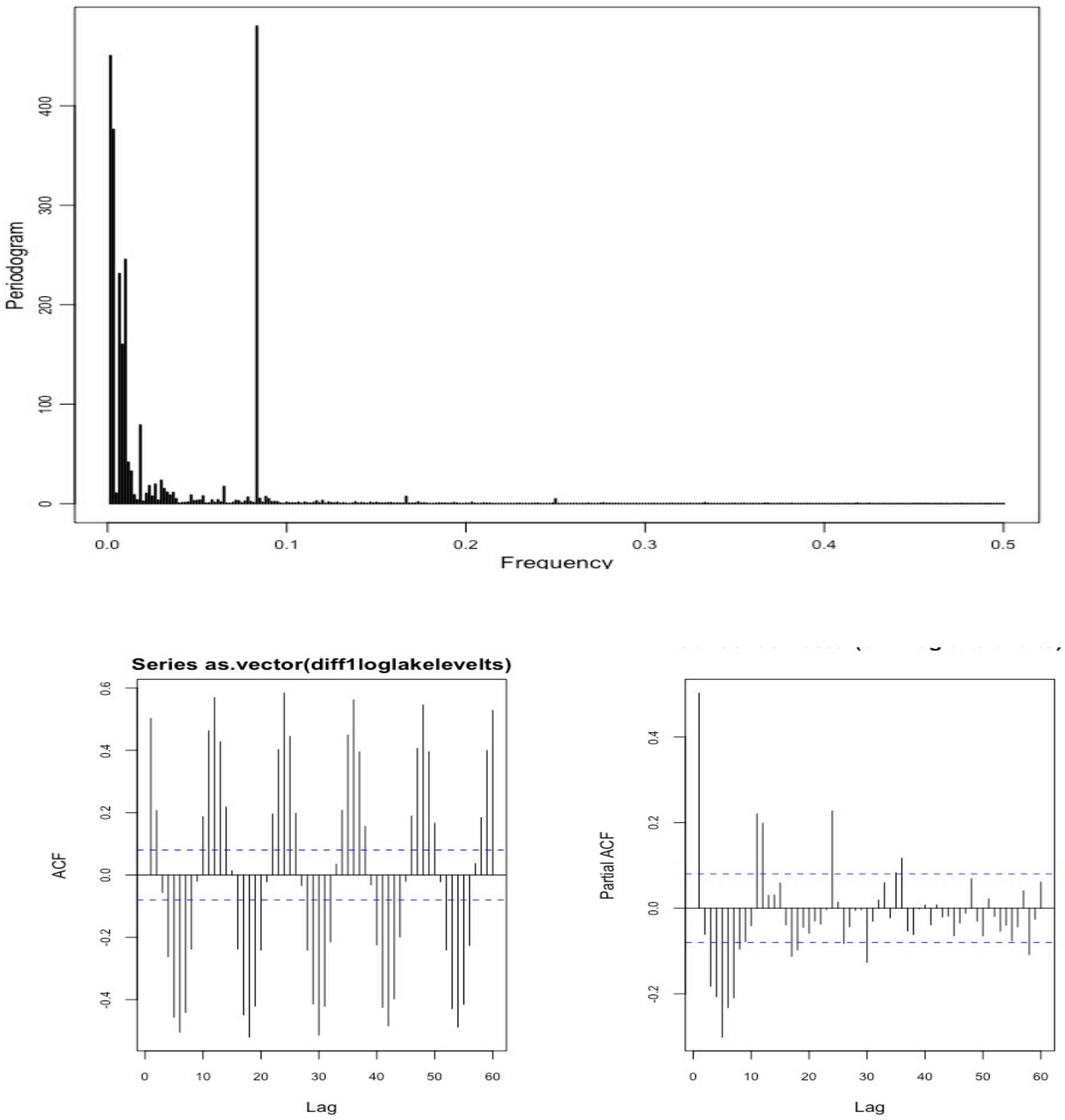

Figure 4: Periodogram, ACF and PACF plots for regular differenced time series

To investigate the periodogram values, top five highest power frequencies were obtained along with their peak values

```
freq     spec
50 0.083333333 480.0320
1  0.001666667 450.4459
2  0.003333333 376.2168
6  0.010000000 245.4883
4  0.006666667 231.1819
> # convert frequency to time periods
> time = 1/top5$f
> time
[1]  12 600 300 100 150
```

The dominant frequency corresponds to a period of  12. That's 12 months, because this is monthly data. We can also seee periodicity at lower frequencies (bigger time periods). That means, we could apply extra MA terms to smoothen the time series such that we get rid of these other low frequency peaks.

### 2.2.4 Seasonal differencing and the resulting plots

As concluded in the previous analysis, we need to apply 12th order differencing in addition to the regular differencing to remove seasonality from the data. Figure 5 shows the resulting ACF and PACF plots of regular and 12th order differenced time series.

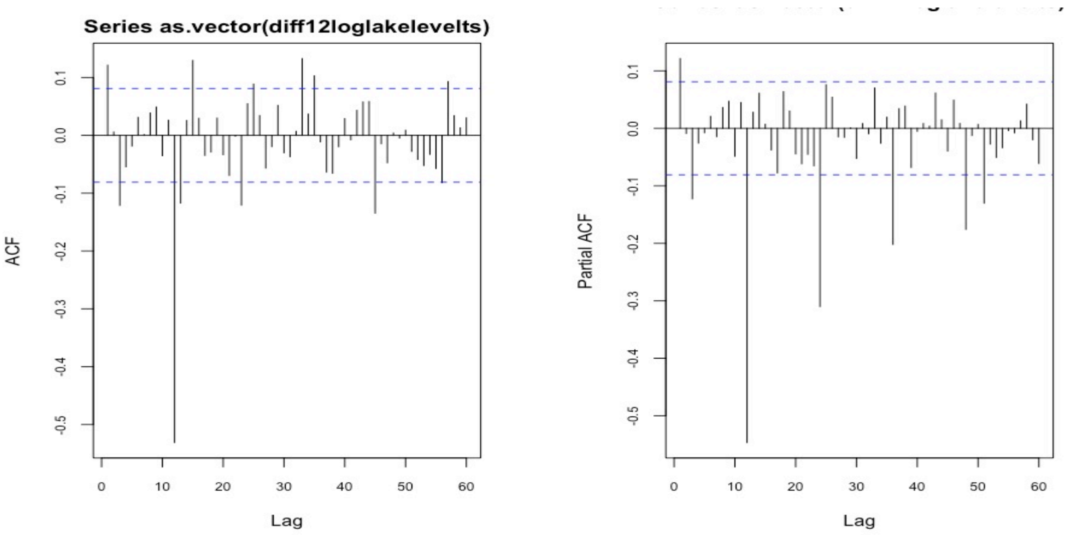

Figure 5: ACF and PACF plots of regular and seasonally differenced time series

Examining the ACF and PACF can be used as a guideline to determine possible models. In Figure 5, we observe that:

- Peaks in ACF at lower lags indicate non-seasonal MA terms, and that the ACF at lag s=12,24 implies that we need to add SMA terms to our model.
- From the PACF, it can be seen that there are no peaks among lags lower than 10 and there are a few peaks---periodically separated---at higher lags. Therefore, we can add SAR terms and test our model.

## 2.3 Model Selection

An important part of the model selection process is getting the p-values of the residuals greater than 5%. This is necessary to ensure that the residuals obtained are uncorrelated. From the previous analysis, it was concluded that adding MA terms and a few SMA terms will help us obtain a model of better fit.

Thus, a few possible models were tested, and the diagnostics were analyzed (see Figure 6) to obtain a model of good fit. The Box-Pierce statistics were significant for all MA models up to MA (12). Hence combinations of MA (12) with SMA models were used to get to a better model.

- ARIMA (0,1,12) x $(1,1,1)_{12}$

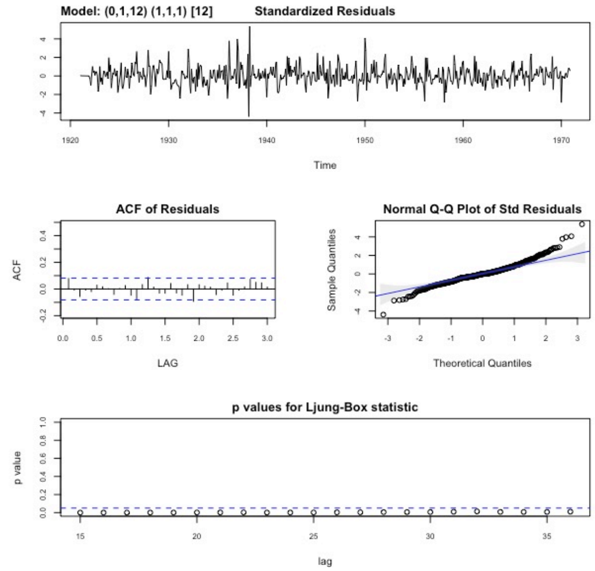

Figure 6: Diagnostic plots of ARIMA (0,1,12) x $(1,1,1)_{12}$

This model is rejected since adding an SAR term made Box-Pierce statistically significant

- ARIMA (0,1,12) x $(1,1,1)_{12}$

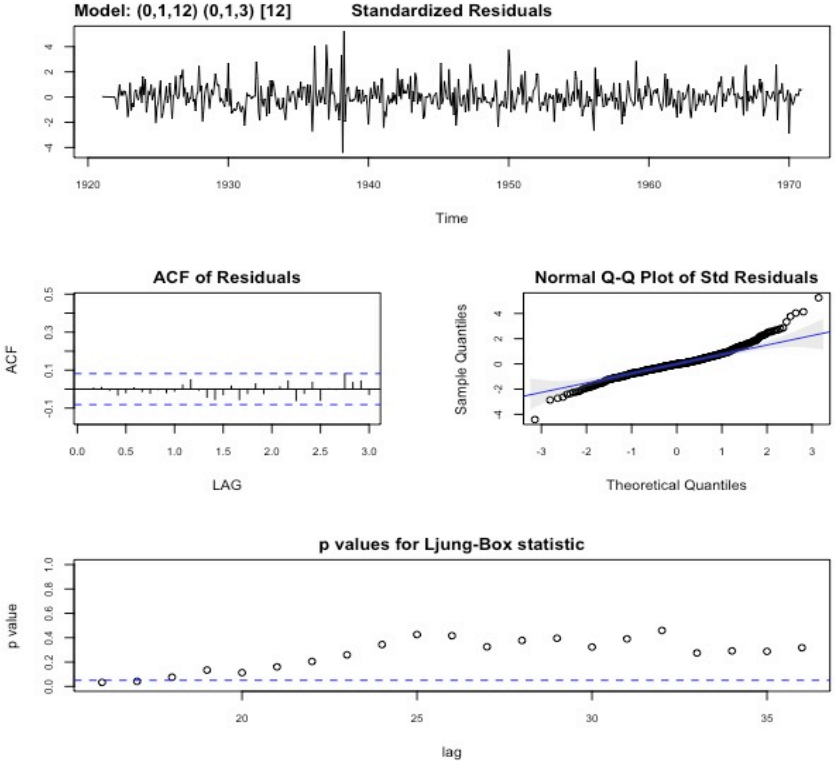

Figure 7: Diagnostic plots of ARIMA (0,1,12) x $(0,1,3)_{12}$

The p-values for Box-Pierce statistics are non-significant. The ACF of residuals look good too. The model meets the assumption that the residuals are independent.

- Using the same diagnostics, several models were tried and tested to obtain a model of best fit. For e.g., ARIMA (0,1,10) x $(0,1,3)_{12}$ , ARIMA (0,1,11) x $(0,1,3)_{12}$ , ARIMA (0,1,12) x $(1,1,3)_{12}$ , ARIMA (0,1,11) x $(1,1,3)_{12}$, etc. All except ARIMA (0,1,11) x $(0,1,3)_{12}$ gave significant Box-Pierce statistics.

- ARIMA (0,1,11) x $(0,1,3)_{12}$ was recognized as a better fitting model for this data. The diagnostics corresponding to this model are shown in Figure 8.

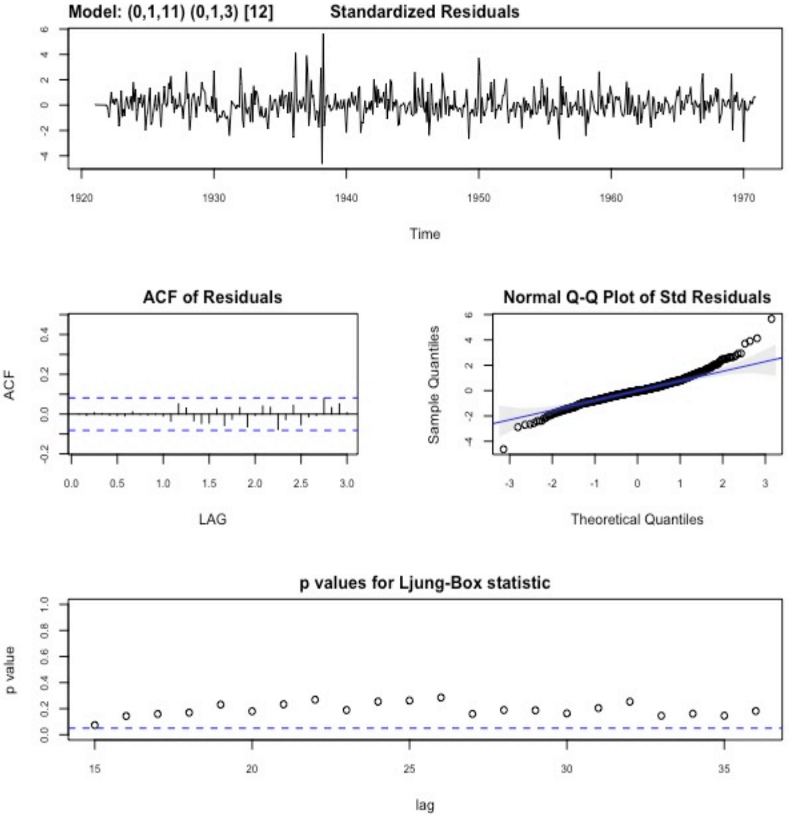

Figure 8: Diagnostic plots of ARIMA (0,1,11) x $(0,1,3)_{12}$

### 2.3.1 Model Coefficients

The model ARIMA (0,1,11) x $(0,1,3)_{12}$ was used to fit on the data. The coefficients output from R was obtained as below

```
Coefficients:
      ma1    ma2     ma3     ma4     ma5    ma6    ma7     ma8
   0.1525 0.0630 -0.1059 -0.0880 -0.0763 0.0372 0.0013 -0.0319
```

```
s.e.  0.0413 0.0422   0.0420   0.0423   0.0423 0.0428 0.0410   0.0431
        ma9     ma10    ma11     sma1    sma2     sma3
     -0.0396 -0.0184 0.0925 -0.9669 0.0585 -0.0389
s.e.   0.0461   0.0459 0.0410   0.0451 0.0598   0.0435

sigma^2 estimated as 0.0007788:  log likelihood = 1253.98, aic = -2477.95

$degrees_of_freedom
[1] 573

$ttable
       Estimate        SE                t.value          p.value
ma1    0.1525        0.0413            3.6896           0.0002
ma2    0.0630        0.0422            1.4908           0.1366
ma3   -0.1059        0.0420           -2.5203           0.0120
ma4   -0.0880        0.0423           -2.0804           0.0379
ma5   -0.0763        0.0423           -1.8048            0.0716
ma6    0.0372        0.0428            0.8695            0.3849
ma7    0.0013        0.0410            0.0309            0.9753
ma8   -0.0319        0.0431             -0.7414           0.4588
ma9   -0.0396        0.0461            -0.8582           0.3911
ma10 -0.0184         0.0459             -0.4002           0.6892
ma11   0.0925        0.0410             2.2535            0.0246
sma1 -0.9669         0.0451            -21.4215          0.0000
sma2   0.0585        0.0598            0.9781            0.3284
sma3 -0.0389         0.0435             -0.8957           0.3708

$AIC
[1] -6.111149
```

```
$AICc
[1] -6.106446

$BIC
[1] -7.008554
```

All the terms except MA1, MA3, MA4, MA11, and SMA1 have p-values greater than 0.05.

The equation for the model can be written as

$(1\text{-}B)\ (1\text{-}B^{12})y_t = (1+\theta B^{11})(1+\Theta B^{12}+\Theta B^{12}+\Theta B^{12})\varepsilon_t$ with $\sigma^2_w = 0.0007788$

## 3. Forecasting

I forecasted the next 12 values using ARIMA (0,1,11) x $(0,1,3)_{12}$ model. The predicted values with 95% confidence interval band are shown in Figure 9.

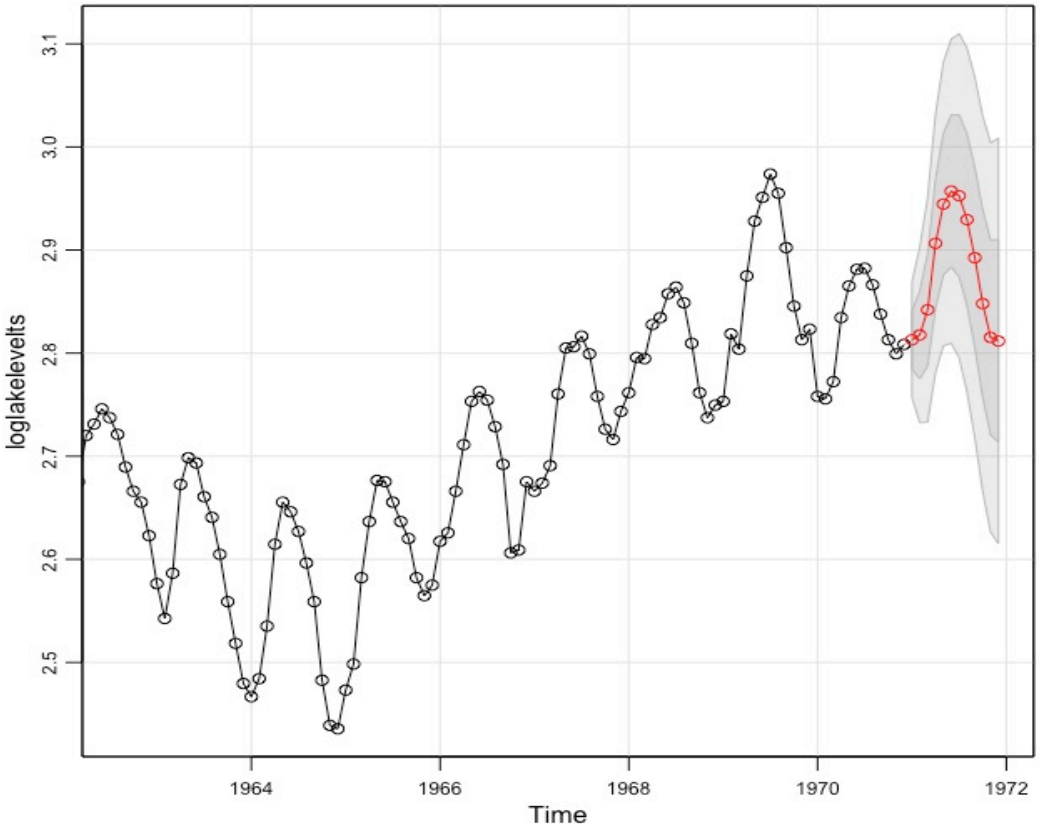

Figure 9: 12-point forecast for ARIMA (0,1,11) x (0,1,3)$_{12}$ model

To test the rigor of my model, I divided the entire dataset in two parts, training data and testing data. Out of 600 points, 588 points were included in the training data and the model ARIMA (0,1,11) x (0,1,3)$_{12}$ was fitted on those training data to get least BIC. Then a prediction of 12 points was made and "predicted" vs "actual" data points were plotted as shown in Figure 10. Note that for a "perfect" prediction, all predicted points must lie on a straight line of zero intercept and unit slope. In reality, through linear regression I obtained a slope of 1.2 and a R-squared value of 0.85. Knowing the limited number of points in the prediction set, and given the closeness of R-squared value to 1.0, it appears that the quality of the prediction is quite reasonable.

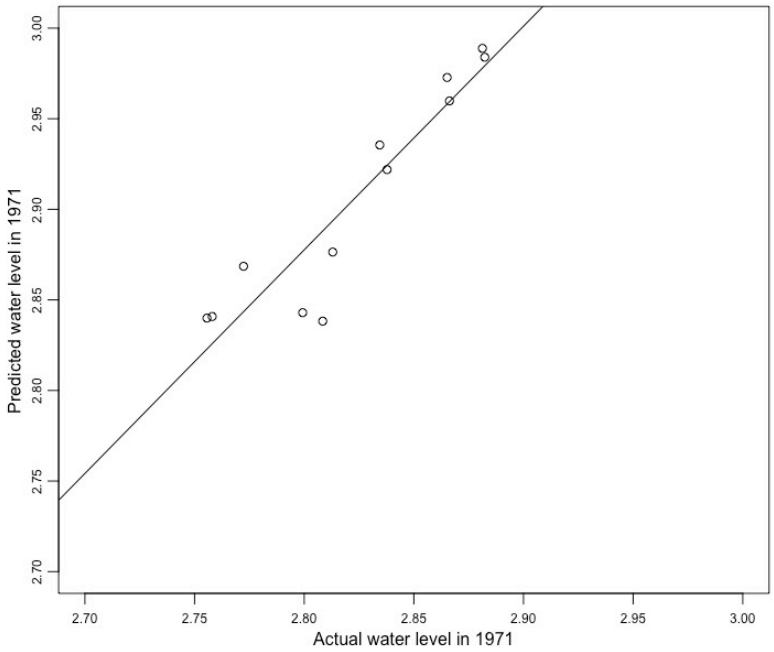

Figure 10: Predicted vs actual 12 points

# 4. Conclusion

For this particular dataset, my final model was an ARIMA (0,1,11) x $(0,1,3)_{12}$.

After taking the regular and seasonal differencing, the ACF and PACF showed sharp peaks at smaller lags suggesting the need for further smoothening, which was addressed by adding more MA terms. To obtain uncorrelated residuals, some trial and error with the precise model parameters was necessary. The auto.arima function suggested ARIMA (1,0,2) x $(2,1,0)_{12}$ which returned worse ACF and PACF plots and correlated residuals. Also, the time series was not stationary, but auto.arima did not suggest any regular differencing. Therefore, I went ahead with ARIMA (0,1,11) x $(0,1,3)_{12}$.

In the previous section I tested how well the model performed with the actual data and the results were satisfactory. But, there may be, are better techniques available to fit this data and I am very interested to learn such techniques, if any.

# 5. References

- Great Lakes Environmental Research Laboratory. (2018.) Great Lakes Water Levels – February 2018. Online PDF: National Oceanic and Atmospheric Administration Great Lakes Environmental Research Laboratory. Retrieved from https://www.glerl.noaa.gov/pubs/brochures/lakelevels/lakelevels.pdf
- Time Series Analysis and its Applications with R examples https://www.stat.pitt.edu/stoffer/tsa4/tsa4.pdf
- Very useful link for time series self-study https://newonlinecourses.science.psu.edu/stat510/node/67/

# 6. R Code

```
set.seed(123)
class(monthly_lake_erie_levels_1921_19)
head(monthly_lake_erie_levels_1921_19, n=5)
#deleting missing values
lakelevel <- na.omit(monthly_lake_erie_levels_1921_19)
#first and last few rows of data
head(lakelevel, n=5)
tail(lakelevel, n=5)
#plot the time series
plot(lakelevel,ylab='Lake Erie Water Levels in Tens of Meters',xlab='Year',type = 'o',pch = 20)
###Exploratory data analysis##
#convert to time series
lakelevelts<-as.ts(read.zoo(lakelevel, FUN = as.yearmon))
head(lakelevelts, n=5)
plot(lakelevelts,ylab='Lake Erie Water Levels in Tens of Meters',xlab='Year',type = 'o',pch = 20)
```

```
lakelevelts
#adf test for stationarity
adf.test(lakelevelts)
#plot acf and pacf
as.vector(lakelevelts)
par(mfrow=c(1,2))
acf(as.vector(lakelevelts),100)
pacf(as.vector(lakelevelts),100)
######log transform data########
loglakelevelts = log(lakelevelts)
#variance stabilization
BoxCox.lambda(lakelevelts)
#qq norm plot
qqnorm(loglakelevelts)
#summary statistics
summary(loglakelevelts)
#qq norm plot
qqnorm(loglakelevelts)
#plot the acf and pacf
par(mfrow=c(1,2))
acf(as.vector(loglakelevelts),60)
pacf(as.vector(loglakelevelts),60)
#decompose the plot
par(mar=c(6,6,2,2),cex.axis=0.75,cex.lab=1)
plot(decompose(loglakelevelts,type = "additive"))
#boxplot for outlier points detection
boxplot(loglakelevelts~cycle(lakelevelts),xlab="months", ylab = "Lake Erie Water Levels in
Tens of Meters" ,main ="Boxplot of Monthly Lake Erie water levels from 1921-1970 ",col
= 'blue',outcol = 'red')
```

```
###1st order difference the data for stationarity,acf and pacf plots###
diff1loglakelevelts = diff(loglakelevelts,1)
par(mfrow=c(1,2))
acf1 <-acf(as.vector(diff1loglakelevelts),60)
pacf1 <-pacf(as.vector(diff1loglakelevelts),60)
#tsdisplay(diff1loglakelevelts)
#Detect seasonality
p = periodogram(diff1loglakelevelts)
dd = data.frame(freq=p$freq, spec=p$spec)
order = dd[order(-dd$spec),]
top5 = head(order,5)
# display the 5 highest "power" frequencies
top5
# convert frequency to time periods
time = 1/top5$f
time
#12th order differencing to remove seasonality
diff12loglakelevelts = diff(diff1loglakelevelts,12)
par(mfrow=c(1,2))
acf12 <-acf(as.vector(diff12loglakelevelts),60)
pacf12 <-pacf(as.vector(diff12loglakelevelts),60)
#tsdisplay(diff12loglakelevelts)
#auto.arima function
auto.arima(loglakelevelts,stepwise = FALSE,approximation=FALSE)
#check diff models
#sarima(loglakelevelts,1,0,0,1,1,0,12)
sarima(loglakelevelts,0,1,11,0,1,3,12)
#sarima(loglakelevelts,0,1,12,1,1,1,12)
#sarima(loglakelevelts,0,1,1,1,1,1,12)
```

```
#sarima(loglakelevelts,0,1,12,0,1,3,12)
#sarima(loglakelevelts,0,1,11,0,1,2,12)
#arima(loglakelevelts,0,1,10,0,1,3,12)
#sarima(loglakelevelts,0,1,12,1,1,3,12)
#sarima(loglakelevelts,0,1,10,0,1,3,12)
#sarima(loglakelevelts,0,1,12,0,1,2,12)
#sarima(loglakelevelts,2,3,0,0,1,1,12)
#sarima(loglakelevelts,12,1,0,0,1,0,12)
#sarima(loglakelevelts,1,1,0,12,1,0,12)
#fit.2 <-sarima(diff12loglakelevelts,1,0,2,2,1,0,12)
##forecast############
for1 = sarima.for(loglakelevelts, 12, 0,1,11, 0,1,3,12)
for1
train_series=loglakelevelts[1:588]
train_series
test_series=loglakelevelts[589:600]
test_series
sarimaModel_1 <-sarima(train_series,0,1,11,0,1,3,12)
forecast =  sarima.for(train_series, 12, 0,1,11, 0,1,3,12)
forecast
plot(loglakelevelts)
par(new=TRUE)
plot(test_series,forecast$pred,
    xlab="Actual water level in 1971",ylab="Predicted water level in 1971",xlim=c(2.7, 3),
ylim=c(2.7, 3))
abline(lm(forecast$pred ~ test_series ))
lm(forecast$pred ~ test_series )
summary(lm(forecast$pred ~ test_series ))
```