SARIMA Model for Colorado River Monthly Flow Data

2022-12-05

Name: Rishab Kulkarni

NetID: rishabk2

Section: 1UG

Abstract

The analyses consists of many components. The first of which involves data exploration to uncover any variance and/or trends present in the time-series data. The second includes in-depth analysis of the A/PACF graphs to confirm whether or not any normal and/or seasonal diff. is required of the time-series data. Upon any differencing, the analyses goes on to draw potential ARIMA models based on the A/PACF graphs of the differenced time-series if applicable. Then, the analyses compares and contrasts the final two proposed ARIMA models and elects the one that best models the time-series data based on outputs from model diagnostics.

Introduction

I am conducting an analyses of time-series data garnered from a point along the Colorado River. My motivation stems from discovering an optimal ARIMA model, which can appropriately model the monthly flow data gathered from the certain point along the river. The reason for the analyses is simple: to examine how and which machine learning models can accurately account for the potential patterns found within the time-series data. What models can describe the variations/trends in the data and provide feasible forecasts for future data points? I hope to answer these queries via the experiment drawn below.

There are multiple goals of the analyses. We must conduct data exploration; this involves examination of trends in the time-series data that may require data transformation through the use of differences. We must then examine the P/ACF of the transformed data to go about drawing up an ARIMA model that models the data and its variations. Yet, we cannot stop here. We must conduct extensive model diagnostics to determine whether or not our proposed ARIMA model is indeed an appropriate fit of the time-series data. Only then can we safely draw conclusions on the ARIMA model we've arrived upon.

The ARIMA model has an integral role for the analyses. ARIMA models can account for the trend and variations in time-series data, along with future values that can be predicted through forecasts. They help us understand the nature of the data and give us an idea of where the data might move. ARIMA models are a useful tool in machine learning and the experiment will demonstrate its unbounded use and applications.

As mentioned before, the time-series data was monthly flow data gathered from a point along the Colorado River. Each measurement corresponds to the flow at that one point along the river. The data was garnered over the course of 600 months or 50 years. 600 data points is enough for us to continue the experiment.

The ARIMA model observes variations/trend across the monthly flow data, factoring in each monthly flow measurement. Thus, we can extract a comprehensive model that properly models the time-series process.

The document shows the 600 observations gathered for each month. In the analyses, we make use of train/test data.

ARIMA #1

The train series does not seem to be stationary, as there are several peaks over the time period. However, there is not much of an apparent trend. We can conduct an Augmented Dickey-Fuller test to corroborate whether or not we must difference the series.

The p-value from the conducted test is much less than 0.01, which suggests that no difference is required. We reject H_0 that the series is not stationary. In essence, we've ample evidence that suggests the train series is stationary with no cause for normal differencing.

Now, we must consider seasonal differencing in the train series. We do this through examination of the A/PACF graphs. The A/PACF graphs provide us insight on potential ARIMA models.

The ACF shows peaks at the lag multiples; Clearly, the graph shows peaks at ~12 and ~24. This indicates that we must use a seasonal difference of the train series.

A/PACF graphs are instrumental for deciding an ARIMA model that can best model the variation in the train series. The experiment demonstrates how analyses of A/PACF graphs can lead to an optimal ARIMA model for a time-series.

We must examine the seasonal component of the ARIMA model. The PACF cuts off at the lag h=1, and the ACF tapers off towards 0. This indicates that p = 1 and q = 0 for the seasonal part of the ARIMA model. The d component = 1 as we use a seasonal difference.

Now, we examine the non-seasonal component. We do so by examining the first few lags for both the ACF and PACF. Again, the PACF cuts out at lag h=1 and the ACF tapers off. This indicates that p = 1 and q = 0 for the non-seasonal component of the ARIMA model. However, now d = 0 as we do not use any normal difference.

The proposed ARIMA model: $(1, 0, 0)$ x $(1, 1, 0)_{12}$. Now, we must conduct model diagnostics to examine the ARIMA model's performance and fit.

ARIMA #1 Diagnostics

Albeit not white noise, the standardized residuals are somewhat stationary. The ACF of residuals are under the dashed blue line. Most of standardized residuals in the Normal Q-Q plot are on the blue line with meager divergence along the tails. Some of the p values for Ljung-Box graph are over the blue line.

More p values must be over the line in the Ljung-Box graph. Also, more std. residuals must be along the dark line in the Normal Q-Q plot. These are shortcomings of the proposed ARIMA model.

The model diagnostics indicate that the ARIMA model is good, but not great. In the experiment, we endeavor to find a better ARIMA model for the train time-series data t1.

ARIMA #2

We can think of the non-seasonal component in another way. Earlier, we set p=1 and q=0 because the ACF tapered towards 0 and the PACF cut out at lag h=1. Now, we can consider the ACF to be cut off at lag h=3, whereas the PACF tapers off towards 0 as shown by the peaks at lag h={12,24,..}. Using this theory, p=0 and q=3 while d remains 1 as we continue with the seasonal difference.

Thus, we propose an ARIMA: $(1, 0, 0)$ x $(0, 1, 3)_{12}$. We just changed the ACF to cut out rather than taper towards 0. Now, we must conduct model diagnostics for the ARIMA to observe its performance in comparison to the previous ARIMA model.

ARIMA #2 Diagnostics

The overall performance of the amended ARIMA is more favorable than that of the original ARIMA model.

The standardized residuals continue to appear stationary, no problem here. The ACF of residuals have more values under the line, an improvement from before. More of the standardized residuals in the Normal Q-Q plot are along the line with lower divergence at the tails than the previous ARIMA model. Far more p values in the Ljung-Box graph are above the line than the prev. ARIMA model, another remarkable improvement. All metrics show improvement in the new ARIMA: $(1,0,0)$ x $(0,1,3)_{12}$.

We can compare another principal metric known as AIC. The lower the AIC, the better the model. The AIC of the original ARIMA $(1,0,0)$ x $(1,1,0)_{12}$ was ~2.6. The AIC of the new ARIMA $(1,0,0)$ x $(0,1,3)_{12}$ is ~2.2, a drastic improvement. Overall, the model diagnostics demonstrate that the new ARIMA $(1,0,0)$ x $(0,1,3)_{12}$ is the preferred model for the data.

As mentioned, we use the train series t1 so we can test the proposed ARIMA model on an unseen batch of data, t2.

The experiment's goal was to find an appropriate ARIMA model that could model the time-series data t1. In the analyses, we found two ARIMA models that account for the variation/trend in t1. Then, we used model diagnostics to decide upon the superior ARIMA model.

The motivation behind the analyses was to find a seasonal ARIMA model that could appropriately model the seasonal time-series data gathered from the monthly flow at a point along the Colorado River. In the analyses, we found two ARIMA models that model the data. The experiment goes further to compare the two ARIMA models using model diagnostics and choose the best one.

Appendix – R Code & Outputs

```
library(astsa)
# loading package for time-series analyses

t <- ts(scan("/Users/rishabkulkarni/Downloads/coloradoflow.dat"))
# scan time-series data

t
```

```
## Time Series:
## Start = 1
## End = 600
## Frequency = 1
##   [1] 0.46 0.53 1.24 1.48 3.69 4.57 2.58 0.69 0.50 1.40 0.54 0.43 0.43 0.40 0.66
##  [16] 1.10 4.34 7.08 3.34 1.27 0.65 0.77 0.66 0.37 0.43 0.38 0.64 2.19 3.68 3.39
##  [31] 1.86 0.61 0.73 0.80 0.66 0.40 0.45 0.49 1.08 1.89 5.49 7.26 3.24 1.29 0.75
##  [46] 1.17 0.65 0.41 0.37 0.48 0.64 1.77 2.83 4.01 1.97 0.57 0.52 0.63 0.50 0.42
##  [61] 0.48 0.52 1.71 2.17 4.35 5.22 2.54 1.97 0.81 1.71 0.61 0.45 0.32 0.43 0.62
##  [76] 1.88 4.53 8.98 5.23 1.39 0.80 0.57 0.55 0.49 0.43 0.46 0.81 1.08 3.08 5.83
##  [91] 2.11 0.67 0.70 0.66 0.61 0.50 0.35 0.38 0.80 1.67 3.61 2.46 1.17 0.66 0.47
## [106] 0.44 0.50 0.48 0.50 0.74 0.84 1.19 7.00 7.95 2.96 1.14 0.50 0.63 0.73 0.50
## [121] 0.50 0.56 1.16 1.10 4.86 9.81 2.72 2.01 0.92 0.52 0.51 0.55 0.42 0.54 1.11
## [136] 1.43 5.49 6.32 1.77 0.87 0.52 0.36 0.49 0.49 0.46 0.42 0.55 1.57 4.45 5.62
## [151] 2.87 1.66 1.08 0.89 0.80 0.52 0.38 0.62 0.62 2.01 3.77 3.85 1.23 0.37 0.30
## [166] 0.44 0.48 0.33 0.32 0.49 0.74 1.59 2.62 2.94 1.77 0.90 1.31 1.23 0.75 0.55
## [181] 0.44 0.43 0.79 1.94 4.14 4.23 1.70 0.66 0.36 0.52 0.41 0.46 0.41 0.47 0.74
```

```
## [196] 1.50 4.76 4.49 2.96 1.14 2.51 1.15 0.90 0.54 0.57 0.57 0.92 1.21 5.35 4.57
## [211] 1.88 0.77 0.43 0.74 0.70 0.42 0.40 0.42 1.14 2.06 5.04 5.93 2.47 2.33 2.02
## [226] 1.14 0.69 0.54 0.36 0.59 0.70 2.10 2.44 3.78 1.31 1.80 0.64 0.67 0.52 0.35
## [241] 0.32 0.45 0.52 0.69 1.42 1.71 0.55 0.30 0.33 0.61 0.43 0.31 0.33 0.68 0.83
## [256] 2.02 4.90 4.46 2.48 1.14 0.60 0.41 0.46 0.33 0.33 0.31 0.61 0.61 1.77 4.90
## [271] 1.37 0.42 0.46 0.46 0.36 0.39 0.37 0.37 0.41 0.56 1.35 0.63 0.16 0.16 0.16
## [286] 0.18 0.22 0.28 0.31 0.33 0.42 0.80 1.73 4.94 1.81 0.64 0.52 0.42 0.39 0.32
## [301] 0.32 0.39 0.56 1.63 4.43 3.27 1.23 1.07 0.67 0.43 0.54 0.39 0.24 0.51 0.86
## [316] 1.86 4.29 2.90 1.63 0.49 0.50 0.56 0.46 0.47 0.39 0.43 0.97 1.99 4.14 0.58
## [331] 2.12 0.62 1.11 0.69 0.58 0.49 0.42 0.36 0.99 1.42 3.07 2.05 0.51 0.27 0.67
## [346] 0.40 0.38 0.34 0.31 0.36 0.53 0.84 2.49 1.83 0.48 0.22 0.50 0.84 0.49 0.44
## [361] 0.43 0.52 0.82 1.34 6.13 4.94 2.05 0.98 0.75 2.22 1.11 0.71 0.50 0.49 0.78
## [376] 3.51 3.96 5.18 1.62 0.56 0.34 0.41 0.45 0.44 0.40 0.41 0.63 1.79 2.66 3.36
## [391] 1.76 0.98 0.55 0.46 0.56 0.48 0.34 0.42 0.63 1.26 4.01 5.10 2.20 0.51 0.28
## [406] 0.42 0.47 0.39 0.40 0.43 0.54 0.93 3.46 3.40 2.06 1.25 0.46 0.62 0.54 0.41
## [421] 0.45 0.39 0.61 1.25 2.13 2.46 0.90 0.59 0.38 0.50 0.58 0.55 0.34 0.44 0.81
## [436] 0.96 3.85 4.04 2.38 1.48 0.72 1.01 0.72 0.57 0.50 0.56 0.80 2.10 4.32 4.12
## [451] 1.21 0.65 0.28 0.41 0.50 0.43 0.42 0.44 0.87 1.61 3.82 5.45 2.64 0.71 0.39
## [466] 0.63 0.58 0.45 0.43 0.49 0.80 1.50 2.42 3.67 1.70 0.52 0.41 0.42 0.43 0.51
## [481] 0.39 0.44 0.51 0.65 2.03 3.56 1.67 0.97 0.51 0.51 0.55 0.41 0.59 0.47 0.54
## [496] 2.79 6.26 6.40 1.94 1.01 0.67 0.45 0.47 0.47 0.48 0.45 0.56 0.65 1.29 3.69
## [511] 1.17 0.81 0.32 0.39 0.51 0.42 0.39 0.42 0.48 0.67 1.57 0.97 0.80 0.39 0.42
## [526] 0.63 0.43 0.34 0.30 0.30 0.72 0.76 1.93 1.95 0.70 0.63 0.28 0.26 0.34 0.40
## [541] 0.46 0.35 0.63 1.11 2.70 3.20 0.68 0.44 0.20 0.23 0.37 0.30 0.35 0.40 0.61
## [556] 1.02 3.17 6.96 4.95 1.98 1.01 0.92 1.04 0.63 0.49 0.66 0.86 1.94 4.92 4.54
## [571] 0.77 0.35 0.39 0.38 0.44 0.45 0.39 0.39 0.42 0.52 1.26 2.26 0.96 0.52 0.30
## [586] 0.62 0.61 0.43 0.37 0.39 0.92 1.98 1.93 2.76 0.80 0.25 0.24 0.42 0.42 0.34
```

```
# display 600 observations from data
```

The above table shows the 600 monthly observations taken from the Colorado River data.

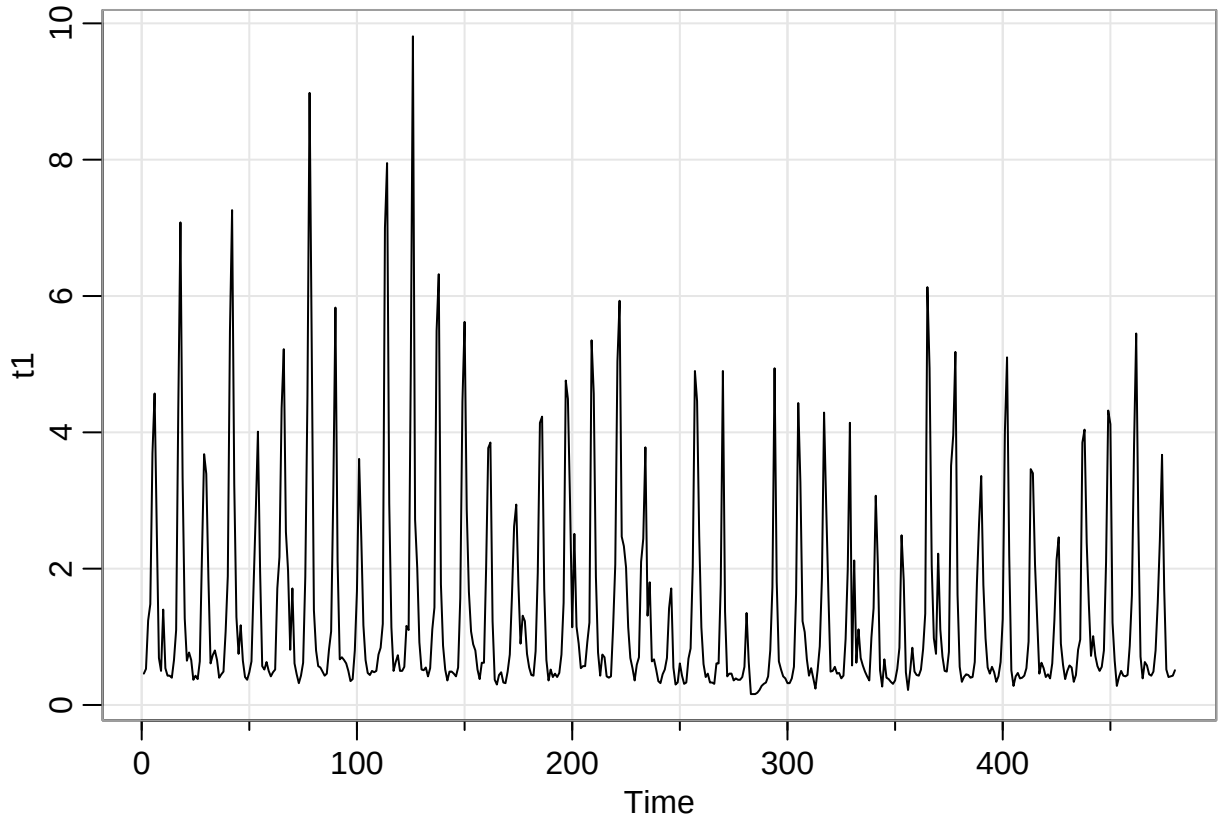
```
library(TSstudio)

s <- ts_split(t, sample.out = 120)  # train-test split


t1 <- s$train # train

t2 <- s$test # test
```

```
tsplot(t1)
```



The above graph depicts the time-series plot of the train time-series t1.

```
library(tseries)
```

```
## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo
```

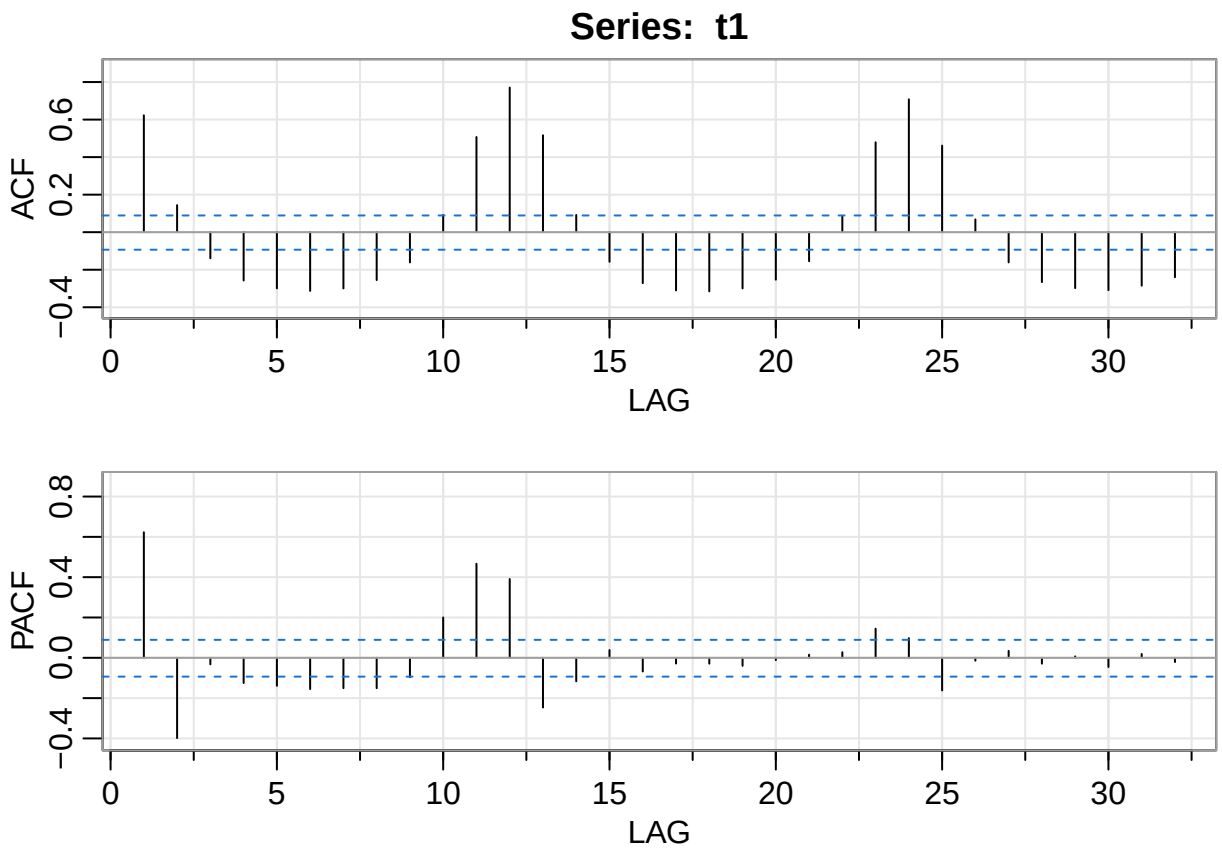
```
adf.test(t1)
```

```
## Warning in adf.test(t1): p-value smaller than printed p-value
```

```
##
##  Augmented Dickey-Fuller Test
##
## data:  t1
## Dickey-Fuller = -13.429, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

The above graphic shows the findings from the Augmented Dickey-Fuller test. We use this test to confirm whether or not we difference the train time-series t1.

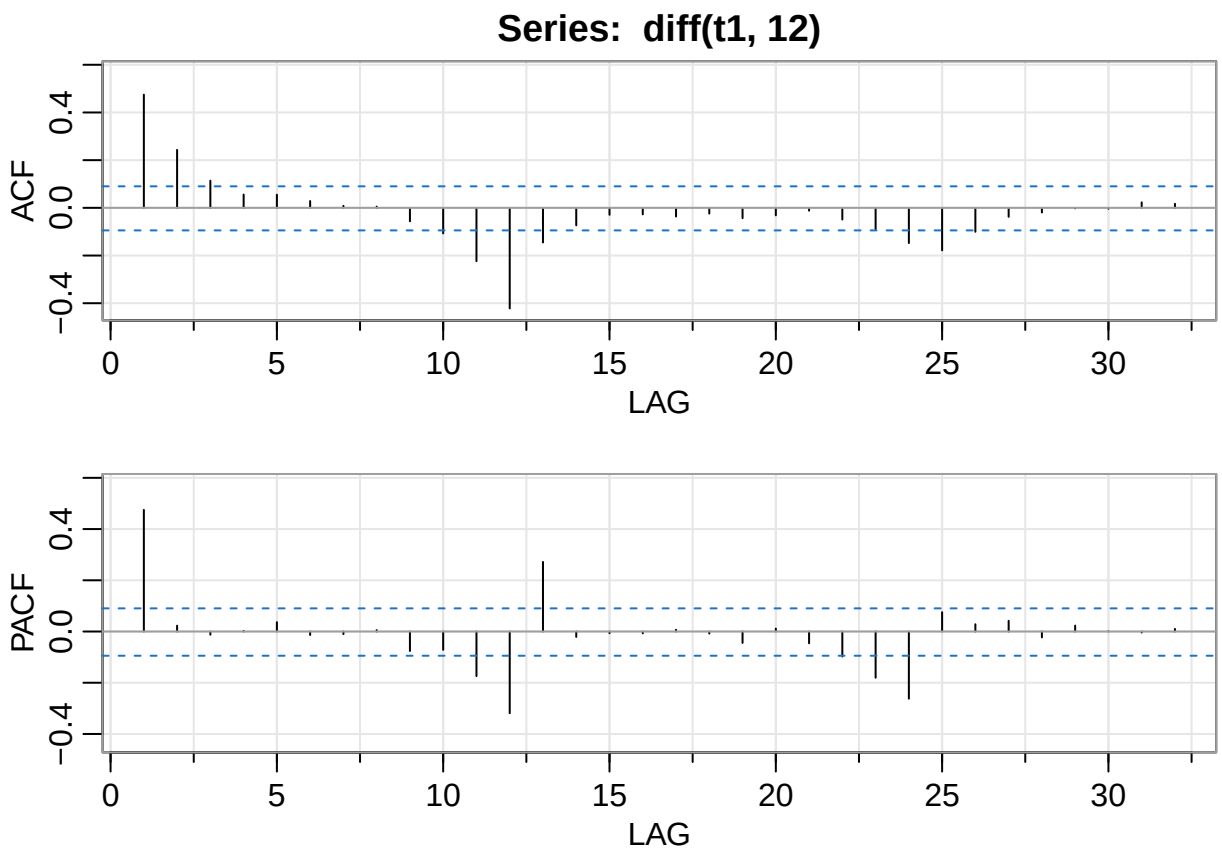
```
acf2(t1)
```




```
##      [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11] [,12]
## ACF  0.62  0.14 -0.14 -0.26 -0.30 -0.31 -0.30 -0.26 -0.16  0.09  0.51  0.77
## PACF 0.62 -0.40 -0.03 -0.13 -0.14 -0.16 -0.15 -0.15 -0.10  0.20  0.47  0.39
##      [,13] [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24]
## ACF   0.52  0.09 -0.16 -0.27 -0.31 -0.32 -0.30 -0.25 -0.15  0.08  0.48  0.71
## PACF -0.25 -0.12  0.04 -0.07 -0.03 -0.03 -0.04 -0.01  0.02  0.03  0.14  0.10
##      [,25] [,26] [,27] [,28] [,29] [,30] [,31] [,32]
## ACF   0.46  0.07 -0.16 -0.27 -0.30 -0.31 -0.28 -0.24
## PACF -0.16 -0.01  0.03 -0.03  0.01 -0.05  0.02 -0.02
```

The above A/PACF graph shows that the train time-series t1 has a seasonal difference. Again, we can conclude this from the peaks at the lag multiples {h=12,24,...}

```
acf2(diff(t1, 12))
```



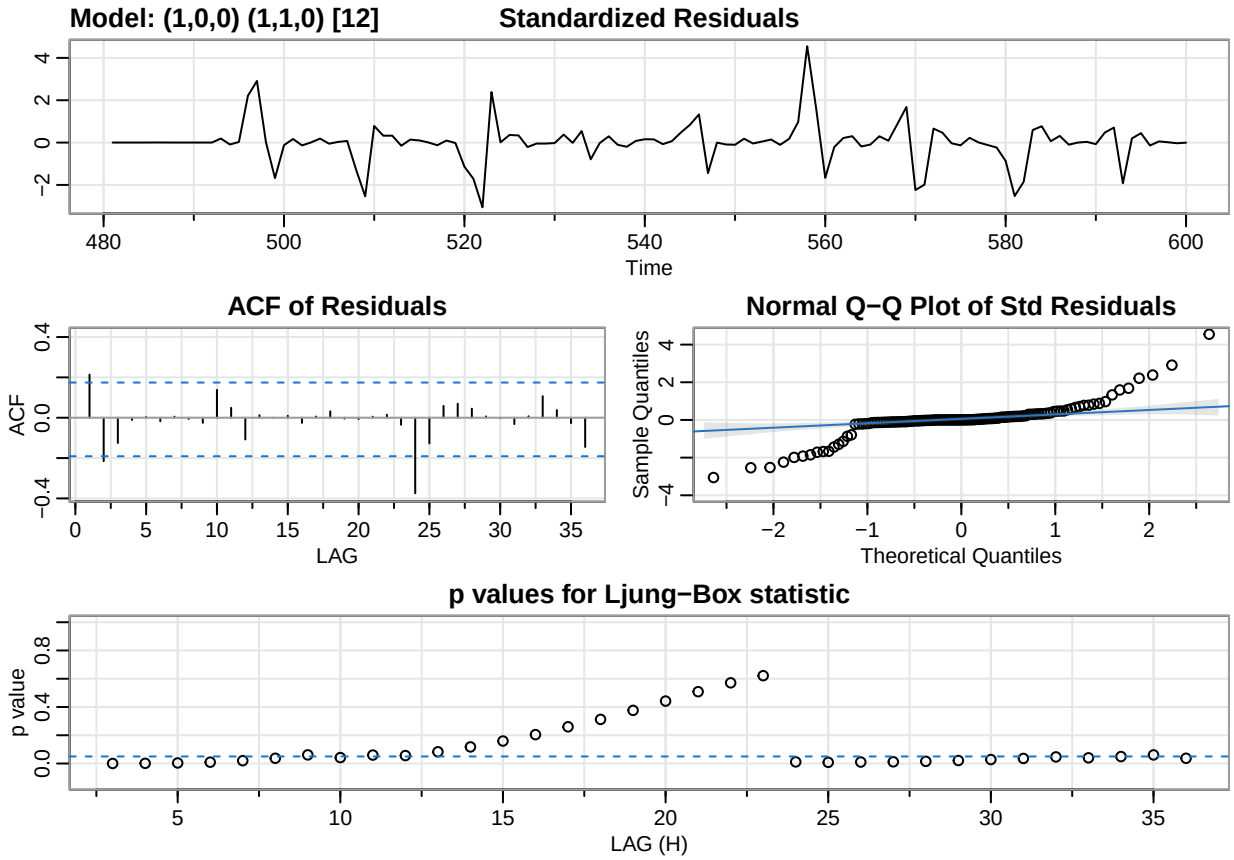

```
##      [,1] [,2]  [,3] [,4] [,5]  [,6]  [,7] [,8]  [,9] [,10] [,11] [,12] [,13]
## ACF  0.47 0.24  0.11 0.06 0.06  0.03  0.01 0.01 -0.06 -0.11 -0.22 -0.42 -0.15
## PACF 0.47 0.02 -0.01 0.00 0.04 -0.01 -0.01 0.01 -0.08 -0.07 -0.17 -0.32  0.27
##      [,14] [,15] [,16] [,17] [,18] [,19] [,20] [,21] [,22] [,23] [,24] [,25]
## ACF  -0.07 -0.03 -0.03 -0.04 -0.02 -0.04 -0.03 -0.01 -0.05 -0.09 -0.15 -0.18
## PACF -0.02 -0.01 -0.01  0.01 -0.01 -0.04  0.01 -0.05 -0.10 -0.18 -0.26  0.08
##      [,26] [,27] [,28] [,29] [,30] [,31] [,32]
## ACF  -0.10 -0.04 -0.02  0.00     0  0.02  0.02
## PACF  0.03  0.04 -0.02  0.02     0  0.00  0.01
```

The above A/PACF provides data on potential ARIMA models that can model the train time-series t1. We observe the A/PACF values at certain peaks to draw up an ARIMA model.

```
model <- sarima(t2, 1, 0, 0, 1, 1, 0, 12) # testing ARIMA model on test series, t2
```

```
## initial  value 0.197057
## iter   2 value -0.189449
## iter   3 value -0.189776
## iter   4 value -0.191608
## iter   5 value -0.191619
## iter   6 value -0.191623
## iter   7 value -0.191623
## iter   8 value -0.191623
## iter   8 value -0.191623
## final  value -0.191623
```

```
## converged
## initial  value -0.154438
## iter   2 value -0.154958
## iter   3 value -0.155051
## iter   4 value -0.155055
## iter   5 value -0.155056
## iter   6 value -0.155056
## iter   6 value -0.155056
## iter   6 value -0.155056
## final  value -0.155056
## converged
```




```
model
```

```
## $fit
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     xreg = constant, transform.pars = trans, fixed = fixed, optim.control = list(trace = trc,
##         REPORT = 1, reltol = tol))
##
## Coefficients:
##          ar1     sar1  constant
##       0.6737  -0.4908   -0.0042
## s.e.  0.0702   0.0834    0.0141
```

```
##
## sigma^2 estimated as 0.7073:  log likelihood = -136.5,  aic = 281
##
## $degrees_of_freedom
## [1] 105
##
## $ttable
##          Estimate     SE t.value p.value
## ar1        0.6737 0.0702  9.5930   0.000
## sar1      -0.4908 0.0834 -5.8836   0.000
## constant  -0.0042 0.0141 -0.2983   0.766
##
## $AIC
## [1] 2.601839
##
## $AICc
## [1] 2.603976
##
## $BIC
## [1] 2.701177
```

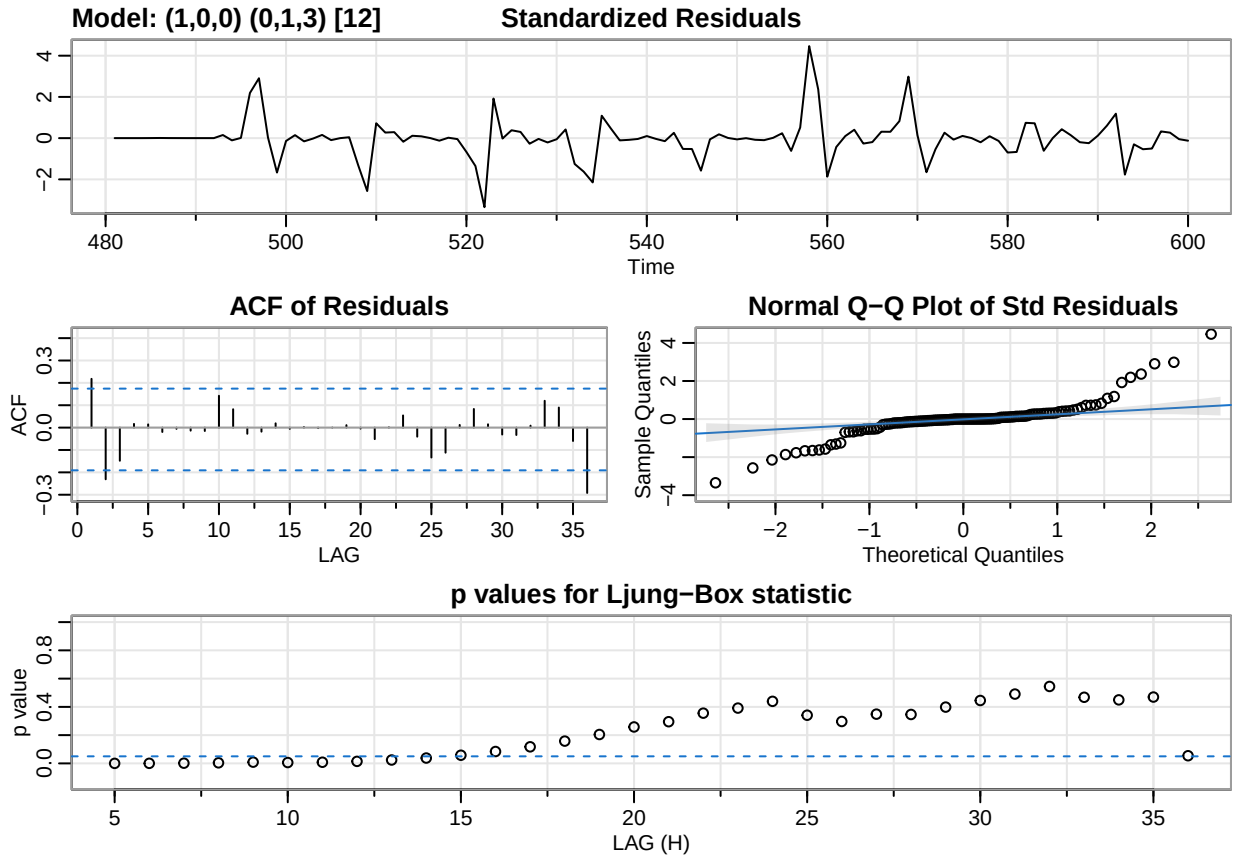
The above graph shows the model diagnostics of the train time-series t1. We observe the data to determine the model's comprehensive performance and fit of t1. Based on these findings, we can elect to settle on this ARIMA model or draw up another one with a better fit of t1.

Now, we fit the new ARIMA model and observe its model diagnostics. We observe if its performance is an improvement from the previous ARIMA model.

```
model2 <- sarima(t2, 1, 0, 0, 0, 1, 3, 12) # testing new ARIMA model on test series, t2
```

```
## initial  value 0.236761
## iter   2 value -0.199446
## iter   3 value -0.263462
## iter   4 value -0.265011
## iter   5 value -0.324610
## iter   6 value -0.354141
## iter   7 value -0.423032
## iter   8 value -0.430031
## iter   9 value -0.431425
## iter  10 value -0.431976
## iter  11 value -0.432240
## iter  12 value -0.432415
## iter  13 value -0.432423
## iter  14 value -0.432423
## iter  15 value -0.432426
## iter  16 value -0.432428
## iter  17 value -0.432428
## iter  18 value -0.432428
## iter  18 value -0.432428
## iter  18 value -0.432428
## final  value -0.432428
## converged
## initial  value -0.368770
## iter   2 value -0.370047
```

```
## iter   3 value -0.371284
## iter   4 value -0.377010
## iter   5 value -0.379069
## iter   6 value -0.380261
## iter   7 value -0.380357
## iter   8 value -0.380366
## iter   9 value -0.380367
## iter  10 value -0.380367
## iter  10 value -0.380367
## iter  10 value -0.380367
## final  value -0.380367
## converged
```




```
model2
```

```
## $fit
##
## Call:
## arima(x = xdata, order = c(p, d, q), seasonal = list(order = c(P, D, Q), period = S),
##     xreg = constant, transform.pars = trans, fixed = fixed, optim.control = list(trace = trc,
##         REPORT = 1, reltol = tol))
##
## Coefficients:
##          ar1     sma1    sma2    sma3  constant
##       0.6646  -1.5231  0.2076  0.4099    0.0005
```

```
## s.e.  0.0718   0.2079  0.1862  0.1098    0.0024
##
## sigma^2 estimated as 0.2648:  log likelihood = -112.17,  aic = 236.33
##
## $degrees_of_freedom
## [1] 103
##
## $ttable
##          Estimate     SE t.value p.value
## ar1        0.6646 0.0718  9.2600  0.0000
## sma1      -1.5231 0.2079 -7.3273  0.0000
## sma2       0.2076 0.1862  1.1151  0.2674
## sma3       0.4099 0.1098  3.7321  0.0003
## constant   0.0005 0.0024  0.1998  0.8420
##
## $AIC
## [1] 2.188255
##
## $AICc
## [1] 2.193702
##
## $BIC
## [1] 2.337262
```

The overall performance of the amended ARIMA is more favorable than that of the original ARIMA model. The experiment uses model diagnostics to compare the performance between two ARIMA models. This allows us to decide on the ARIMA model that fits our time-series data the best.