Time-Series Analysis Project

Novin Shahroudi

December 2018

1 First time series

The first dataset for my experiment was "WL45.txt". This dataset is from United States Geology Survey containing 993 monthly observations. The observation begins from Jan 1930 to Sep 2012. Each observation is the mean value of the surface water at given site and location.

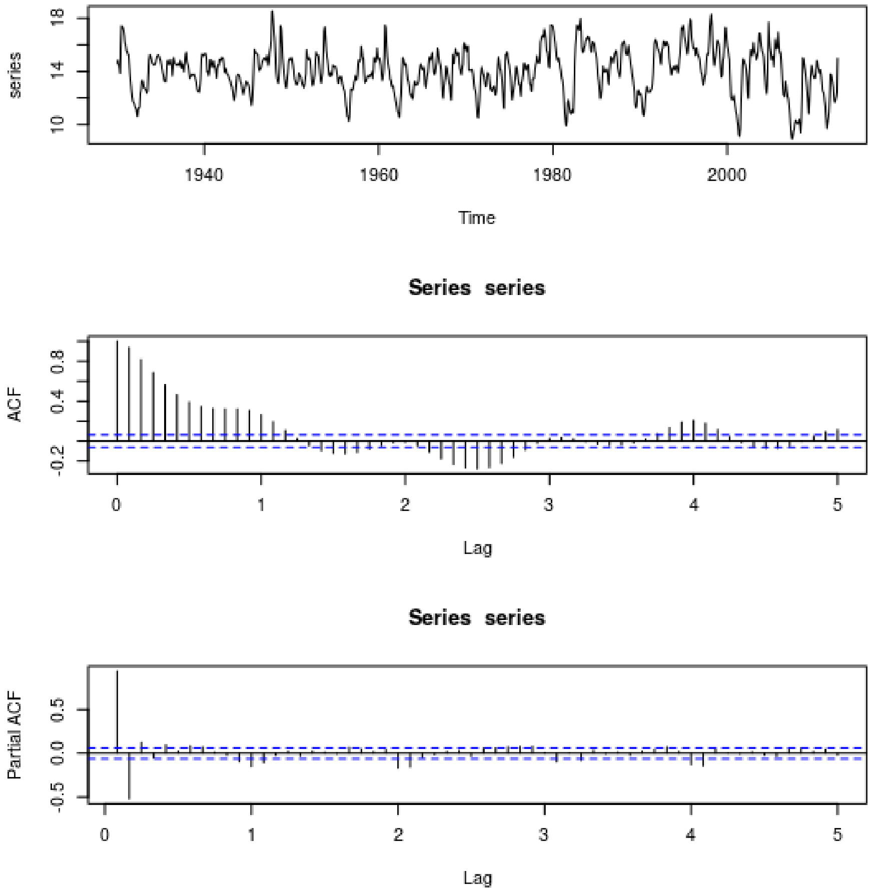


Figure 1: WL45 alongside the ACF and PACF plots

In Fig. 1 we can see that there is no global trend in the data and possibly no local trend. Dickey-Fuller and Phillips-Perron tests confirmed the series is stationary. Despite this, when looking at the ACF it seems that there is some seasonality however weak. Also the correlation decreases slowly which can be an indicator of non-stationarity. To this end, both seasonal and non-seasonal ARIMA models tested.

```
Dickey-Fuller = -6.841, Truncation lag parameter = 7, p-value = 0.01
```

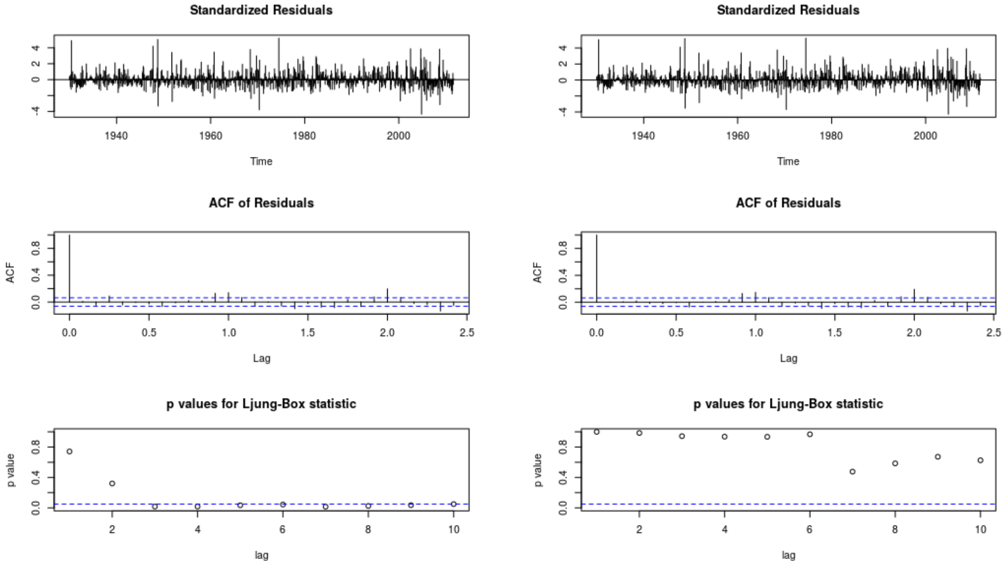
Listing 1: Phillips-Perron Unit Root Test

```
Dickey-Fuller = -5.7226, Lag order = 9, p-value = 0.01
alternative hypothesis: stationary
```

Listing 2: Augmented Dickey-Fuller Test

1.1 (S)ARIMA family

Based on Fig. 1 and PACF we start with ARIMA model with AR(3) as 3 lags are above the limits. Result of this model shown in Fig. 2-(a). The Ljung-Box test is showing high correlation between residuals which means the model is not appropriate. Again based on study of the ACF of the original series or ACF of the obtained residuals We can conclude for at least two MA terms. I chose MA(3) although the ACF suggest even more. This time the residuals are not correlated and seems to be a promising model, however still some peaks above the limits can be seen in the ACF in Fig. 2-(b).

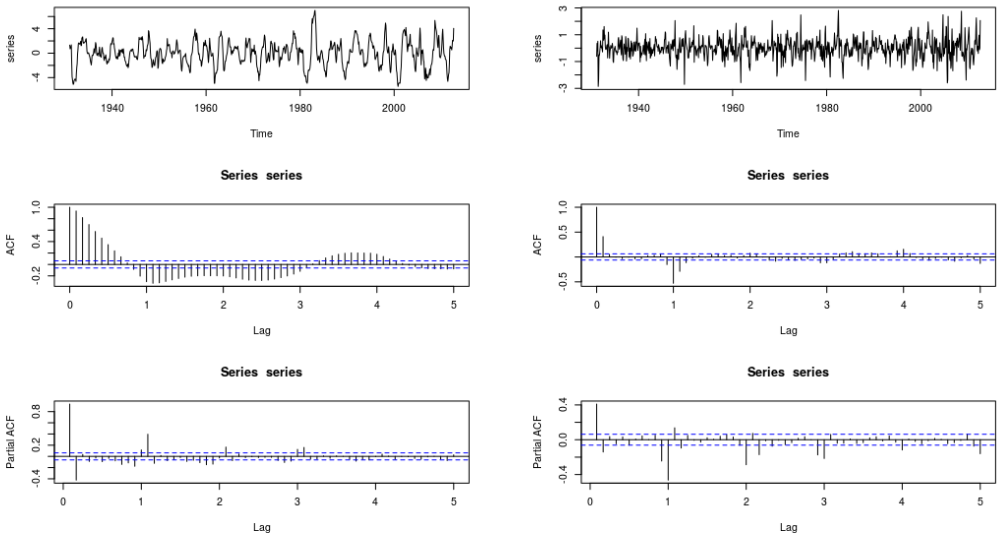


(a) `ARIMA(3,0,0)`

(b) `ARIMA(3,0,3)`

Figure 2: Diagnosis of the fitted ARIMA models

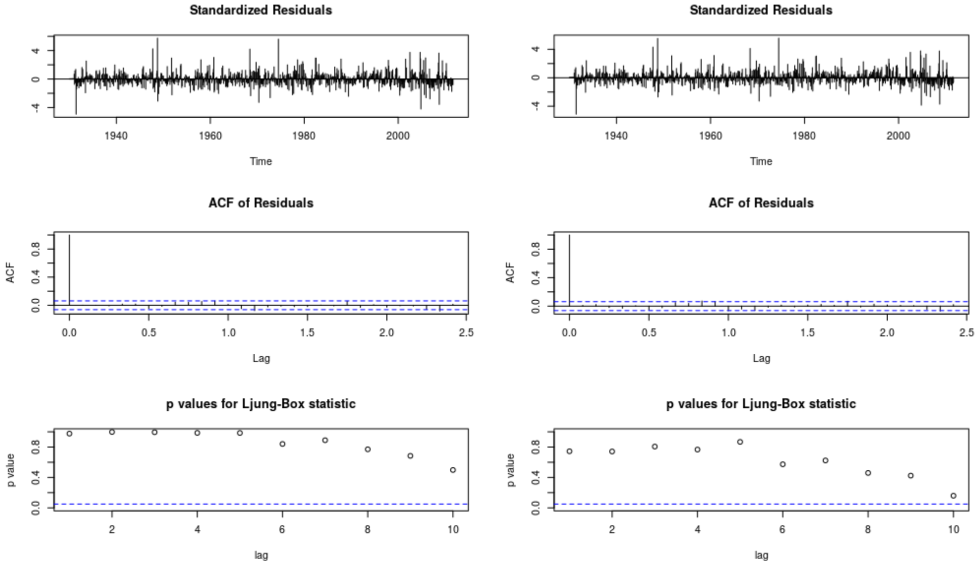
As stated before some seasonal non-stationarity might exist and hence seasonal difference may help to remove it. Fig. 3-(a) shows the seasonal difference with `period=12`. As depicted in Fig. 3-(b) an ordinary difference is also applied together with the seasonal one for a test, but results show that it would be an overkill due to negation of the correlations so we will stick with the seasonal difference alone.



(a) Seasonal Difference (b) Seasonal and Ordinary Difference

Figure 3: Result of making the series stationary with difference differences

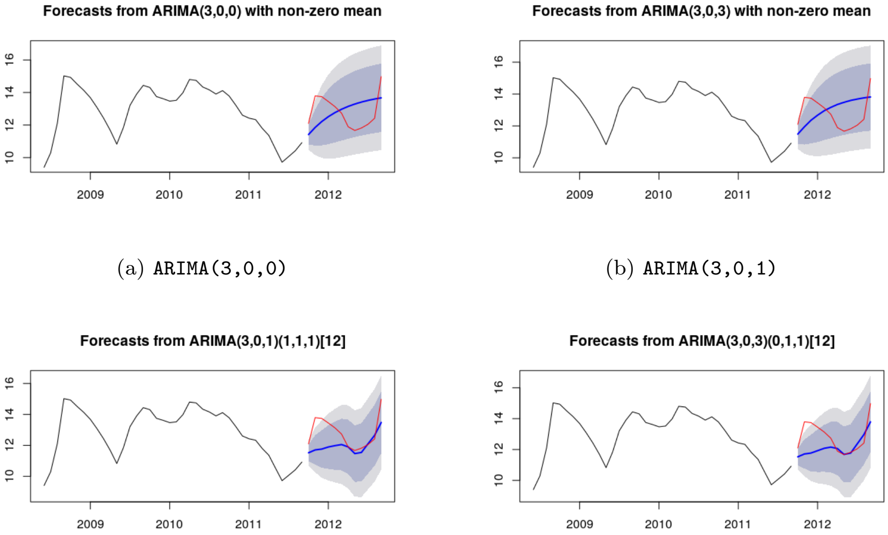
As it can be seen from Fig. 3-(a) PACF there could be up to 3 non-seasonal AR terms and up to 8 non-seasonal MA terms. Similarly, we can have at least 1 term for seasonal AR and 1 term for seasonal MA. Consequently, I used two different seasonal-ARIMA models and their diagnosis can be seen in Fig. 4.One sidenote about why SARIMA(3,0,1)x(0,1,1) instead of SARIMA(3,0,3)x(1,1,0) is that the MA term (based on ACF plot) shows more significance and it makes more sense to go with this one.



(a) `SARIMA(3,0,1)x(1,1,1)_12` (b) `SARIMA(3,0,3)x(0,1,1)_12`

Figure 4: Diagnosis of the fitted SARIMA models

Point forecast and prediction intervals of these models depicted in Fig. 5 with 0.95 and 0.85 confidence levels each colored as light and dark blue respectively. The test data colored as red.



(a) `ARIMA(3,0,0)`

(b) `ARIMA(3,0,1)`

(c) `SARIMA(3,0,1)x(1,1,1)_12`

(d) `SARIMA(3,0,3)x(0,1,1)_12`

Figure 5: Forecast of the ARIMA/SARIMA fitted models on WL45

	MAE	MSE	MAPE
ARIMA(3,0,0)	1.176005	1.278660	9.229816
ARIMA(3,0,1)	1.201571	1.320867	9.502182
SARIMA(3,0,1)x(1,1,1)	0.857519	1.126478	6.362497
SARIMA(3,0,3)x(0,1,1)	0.8378983	1.0854739	6.2424951

Table 1: Performance of each (S)ARIMA model based on different error measurements

Looking at the error measurements suggests that the second SARIMA model performs better but according to the Akaike criteria, it suggests that `SARIMA`$(3,0,1)$`x`$(1,1,1)_{12}$ is the best model. Outputs of the two seasonal models is as follows:

```
Coefficients:
        ar1 ar2 ar3 ma1 sar1 sma1
     0.7938 0.2630 -0.1468 0.6323 -0.1076 -0.9527
s.e. 0.1568 0.2263 0.0855 0.1495 0.0349 0.0166

sigma^2 estimated as 0.1792: log likelihood = -558.76, aic = 1131.52
```

Listing 3: `SARIMA`$(3,0,1)$`x`$(1,1,1)_{12}$ model output

```
Coefficients:
        ar1 ar2 ar3 ma1 ma2 ma3 sma1
     0.8754 0.9461 -0.8473 0.5354 -0.8383 -0.3881 -0.9626
s.e. 0.0837 0.1328 0.0755 0.0936 0.0853 0.0614 0.0160

sigma^2 estimated as 0.1799: log likelihood = -561.44, aic = 1138.88
```

Listing 4: `SARIMA`$(3,0,3)$`x`$(0,1,1)_{12}$ model output

To this end, among the models that I tried, I picked SARIMA$(3,0,1)$x$(1,1,1)_{12}$ as the best model. The model formulation is as follows. [1]

$$\Phi(B^m)\phi(B)\nabla_m^D\nabla^d Z_t = \Theta(B^m)\theta(B)A_t \tag{1}$$

$$\begin{aligned}\Phi(B^{12})\phi(B)\nabla_{12}^1\nabla^0 Z_t &= \Theta(B^{12})\theta(B)A_t\\ (1-\Phi_1B^{12})(1-\phi_1B-\phi_2B^2-\phi_3B^3)(Z_t-Z_{t-12}) &= (1+\Theta_1B^{12})(1+\theta_1B)A_t\end{aligned}$$

NB: Due to lack of time and make the formulas shorter I only considered the first non-seasonal AR term but actually there should be 3 terms!

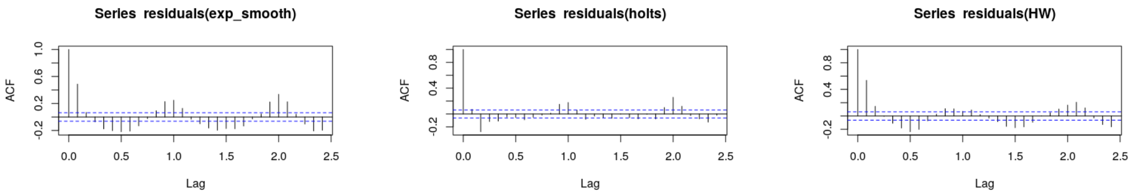
$$\begin{gathered}(1-\Phi_1B^{12})(1-\phi_1B)(Z_t-Z_{t-12}) = (1+\Theta_1B^{12})(1+\theta_1B)A_t\\ (1-\phi_1B-\Phi_1B^{12}+\Phi_1\phi_1B^{13})(Z_t-Z_{t-12}) = (1+\theta_1B+\Theta_1B^{12}+\Theta_1\theta_1B^{13})A_t\end{gathered}$$

$$\begin{aligned}Z_t = {} & \phi_1Z_{t-1}+\Phi_1Z_{t-12}-\Phi_1\phi_1Z_{t-13}+Z_{t-12}-\phi_1Z_{t-13}-\Phi_1Z_{t-24}-\Phi_1\phi_1Z_{t-25}\\ & +A_t+\theta_1A_{t-1}+\Theta_1A_{t-12}+\Theta_1\theta_1A_{t-13}\end{aligned}$$

$$\boxed{\begin{aligned}Z_t = {} & (0.7938)Z_{t-1}+(-0.1076)Z_{t-12}-(-0.1076\times 0.7938)Z_{t-13}+Z_{t-12}\\ & -(0.7938)Z_{t-13}-(-0.1076)Z_{t-24}-(-0.1076\times 0.7938)Z_{t-25}\\ & +A_t+(0.6323)A_{t-1}+(-0.9527)A_{t-12}+(-0.9527\times 0.6323)A_{t-13}\end{aligned}}$$

1.2 HoltWinter family

Based on the analysis performed in the previous step we know that some seasonality terms exist in the series which makes the Holt-winter model to be more appropriate. However, different variations of the holt-winter, meaning that exponential smoothing and holt's method were used to make a comparison. Based on Fig. 6 non of these models are appropriate since the residuals are correlated. This is also suggested by the outcome of the models by studying the errors (Table 2) and the forecast results in Fig. 7.



(a) Exponential Smoothing

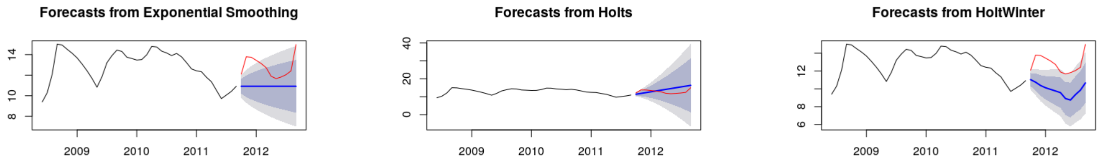
(b) Holts

(c) Holts Winter

Figure 6: Residuals of the HW-family models studied with auto-correlation function

As depicted in Fig. 7-(c) it seems that the Holt-winter model was able to capture the overall pattern of the time series but the level is not correct. Applying different transformation approaches such as log-transform and difference did not make much difference and the final result was more or less the same.

[1] Notation from https://stats.stackexchange.com/questions/82197/how-to-write-seasonal-arima-model-mathematically used.



(a) `Exponential Smoothing` (b) `Holt's Method` (c) `Holt-Winters`

Figure 7: Forecast of the HW-family models

	MAE	MSE	MAPE
Exponential Smoothing	1.894033	2.130382	14.307822
Holt's Method	1.854273	2.163968	14.874172
Holt-Winters	2.891893	2.983005	22.401574

Table 2: Performance of each HW-family based on different error measurements

To this end, the Holt's method was performing the best among the HW-family and I pick this model. Output of the model in R is as follows:

```
Smoothing parameters:
 alpha: 0.8415971
 beta : 0.005600746
 gamma: 1
```

Listing 5: Holt-Winters exponential smoothing with trend and additive seasonal component

$$\hat{Z}_{t+p|t} = a_t + b_t p \tag{2}$$

$$\begin{aligned} a_t &= \alpha z_t + (1-\alpha)\hat{z} = \alpha z_t + (1-\alpha)(a_{t-1} + b_{t-1}), \\ b_t &= \beta(a_t - a_{t-1}) + (1-\beta)b_{t-1} \end{aligned} \tag{3}$$

$$\hat{Z}_{t+p|t} = a_t + b_t p$$

$$\begin{aligned} a_t &= 0.841 z_t + (1-0.841)(a_{t-1} + b_{t-1}), \\ b_t &= 0.005(a_t - a_{t-1}) + (1-0.005)b_{t-1} \end{aligned}$$

1.3 Conclusion of analyzing time series WL45

The best model from (S)ARIMA family was $\texttt{SARIMA}(3,0,1)\texttt{x}(1,1,1)_{12}$ as it gained the minimum errors among all other and had a better Akaike compared to the other SARIMA model. Also from the second part we can conclude that the Holt's method was the best model yet not a reliable as the residuals from the fitted models were not independent and uncorrelated. In general the SARIMA model was the best. Fig. 8 demonstrates the forecast of the best models found.

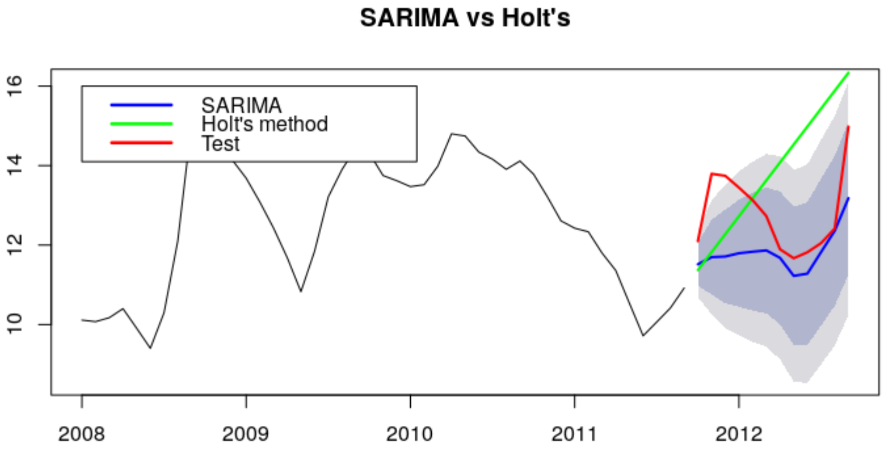


Figure 8: Forecast of the best models for the first series

2 Second time series

The second time series that needed to be analyzed was "TC093s". This time series represent monthly observations for the traffic accidents with casualties. The total number of observations were 336 ranging from Jan 1990 to Dec 2017. As we can see in the Fig. 9 the series is not stationary and contains global trend as well as seasonality. This is also obvious from the (P)-ACF plot(s).

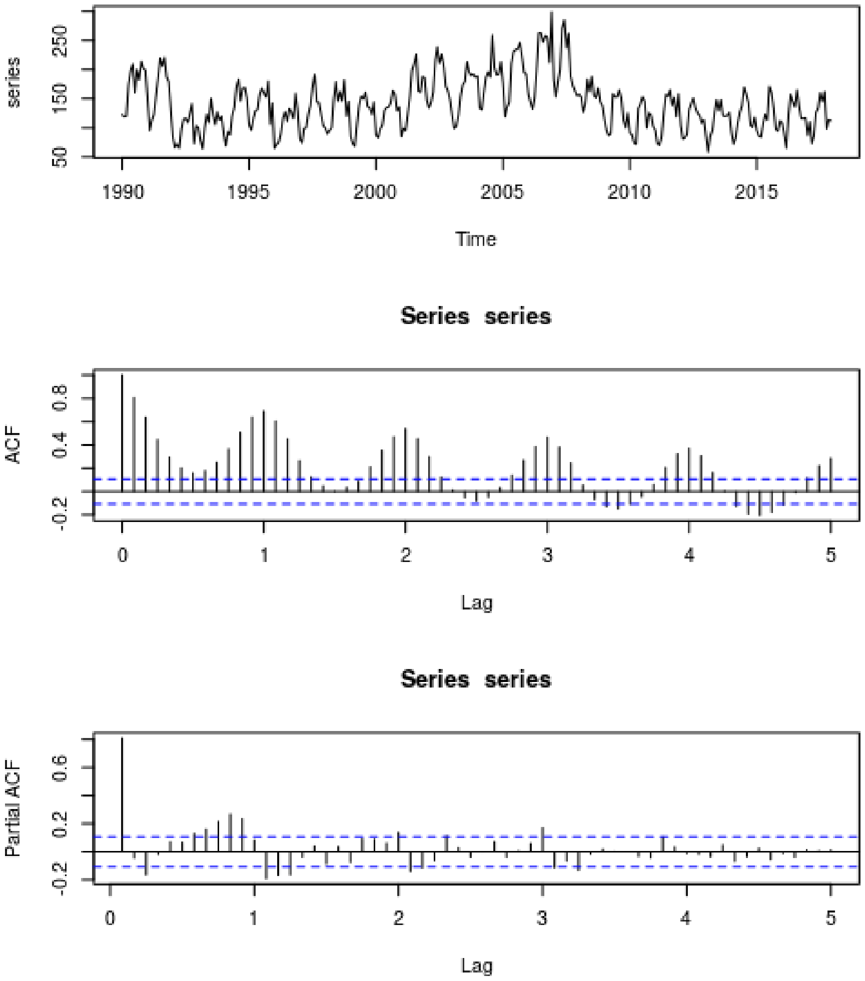


Figure 9: TC093s.txt plots

Similar to the previous series stationary tests run for this series. Although it looks like some local trends exist but the test did not reveal any non-stationarity, however, from the ACF we can see that there is a strong seasonality. The test outcomes are as follows:

```
Dickey-Fuller = -6.1154, Truncation lag parameter = 5, p-value = 0.01
```

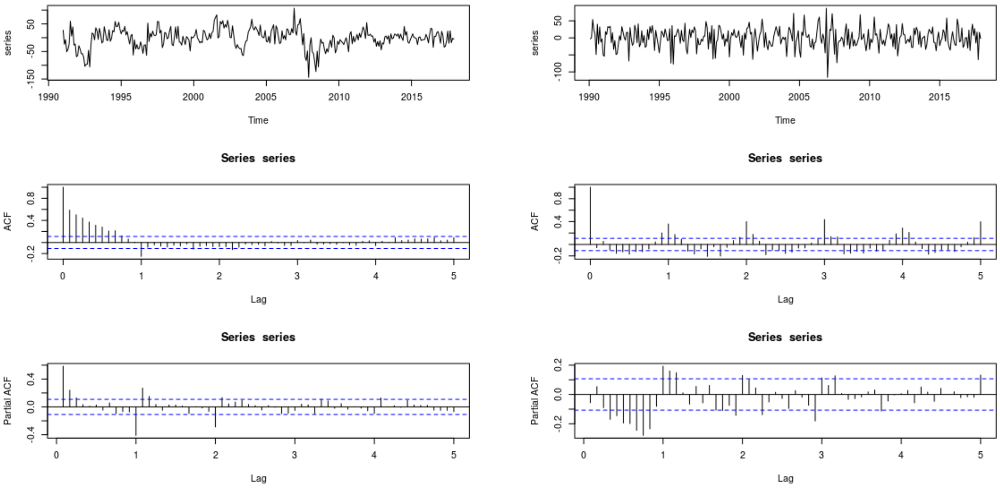
Listing 6: Phillips-Perron Unit Root Test

```
Dickey-Fuller = -4.1344, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

Listing 7: Augmented Dickey-Fuller Test

2.1 (S)ARIMA family

Looking at the ordinary difference and seasonal difference (`period=12`) as depicted in Fig. 10 shows that the ordinary differencing is an overkill and the seasonal differencing is better stationarizes the series.

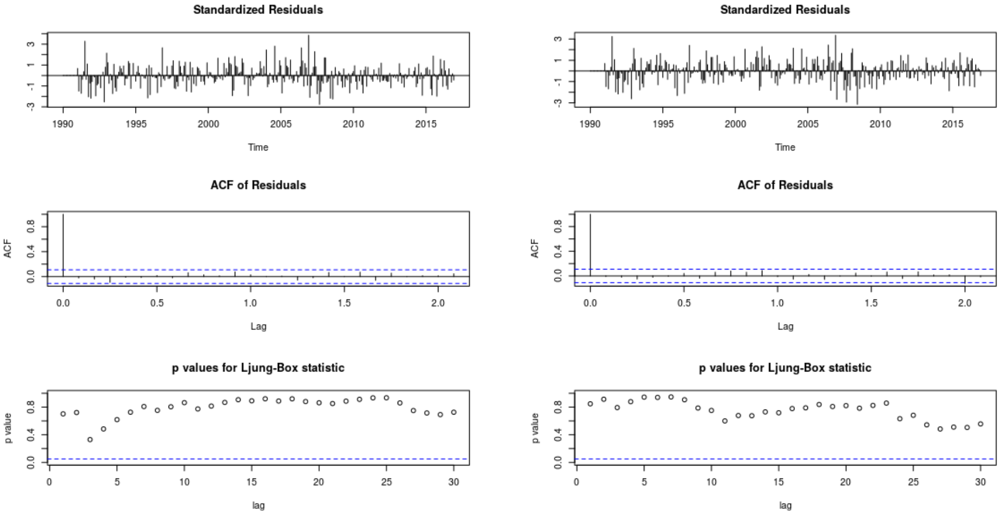


(a) Seasonal Difference

(b) Seasonal and Ordinary Difference

Figure 10: Result of making the series stationary with difference differences

Looking at the Fig. 10-(a) we can see that there are at least 2 seasonal-AR terms and possibly one or more seasonal-MA term(s). Also, from the PCA can be seen that there at most 3 non-seasonal AR terms. To this end, two different SARIMA models as depicted in Fig. 11 fitted for this data.

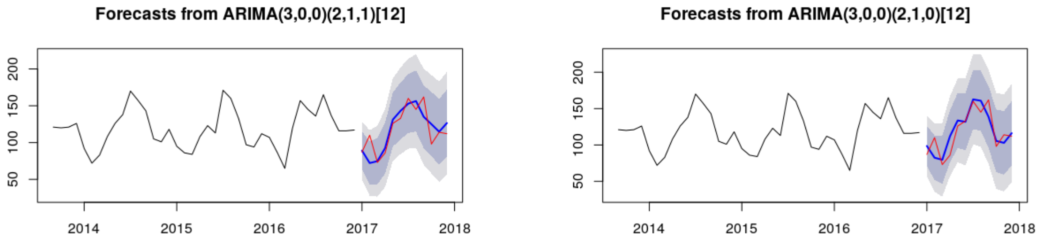


(a) `SARIMA(3,0,0)x(2,1,1)`

(b) `SARIMA(3,0,0)x(2,1,0)`

Figure 11: Diagnosis of the fitted SARIMA models

Results obtained from each model depicted in Fig. 12 for the forecasting and the errors in Table 3.



(a) `SARIMA(3,0,0)x(2,1,1)`

(b) `SARIMA(3,0,0)x(2,1,0)`

Figure 12: Forecast of the fitted SARIMA fitted models on TC093s

	MAE	MSE	MAPE
SARIMA(3,0,0)x(2,1,1)	12.59178	16.96720	10.70865
SARIMA(3,0,0)x(2,1,0)	12.11275	14.93572	11.07459
ARIMAX(3,0,0)x(2,1,1)	18.51041	20.80500	16.46115

Table 3: Performance of each (S)ARIMA(X) models based on different error measurements

Although `SARIMA(3,0,0)x(2,1,1)` is not performing better than `SARIMA(3,0,0)x(2,1,0)` on the error measures but based on Akaike criteria it is a better model:

```
Coefficients:
         ar1 ar2 ar3 sar1 sar2 sma1
      0.5196 0.2069 0.1890 -0.1035 -0.0254 -0.8525
s.e. 0.0563 0.0621 0.0559 0.0806 0.0761 0.0618

sigma^2 estimated as 403.6: log likelihood = -1387.96, aic = 2789.92
```

Listing 8: `SARIMA(3,0,0)x(2,1,1)` model output

```
Coefficients:
         ar1 ar2 ar3 sar1 sar2
      0.4931 0.2130 0.1664 -0.7492 -0.3965
s.e. 0.0559 0.0614 0.0558 0.0554 0.0552

sigma^2 estimated as 460.8: log likelihood = -1403.85, aic = 2819.7
```

Listing 9: `SARIMA(3,0,0)x(2,1,0)` model output

As a conclusion the `SARIMA(3,0,0)x(2,1,1)` mathematical formulation would be as follows:

$$\Phi(B^m)\phi(B)\nabla_m^D\nabla^d Z_t = \Theta(B^m)\theta(B)A_t \tag{4}$$

$$\begin{aligned}\Phi(B^{12})\phi(B)\nabla_{12}^1\nabla^0 Z_t &= \Theta(B^{12})\theta(B)A_t\\(1-\Phi_1 B^{12}-\Phi_2 B^{24})(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)(Z_t-Z_{t-12}) &= (1+\Theta_1 B^{12})A_t\end{aligned}$$

NB: Due to lack of time I did not simplify and expand the equations!

$$(1+0.1035B^{12}+0.0254B^{24})(1-0.5196B-0.2069B^2-0.1890B^3)(Z_t-Z_{t-12}) = (1-0.8525B^{12})A_t$$

2.2 ARIMAX

Among the available time series in the second directory there was only one dataset which matched the dates of the first time series that I was supposed to analyze (`TC093s.txt`) and that was `TC1422s.csv`. This time series was about freight traffic on railways and also a monthly data. From now I refer to the first series as Z and second series as X. The X had 1 year of difference in starting and ending with the Z, in another words, Z started from Jan 1990 and ending Dec 2017 while Z started from Jan 1991 and ending Feb 2018. In order to align the starting and ending dates of both I dropped the first 12 months of the Z and last 2 months of the X.

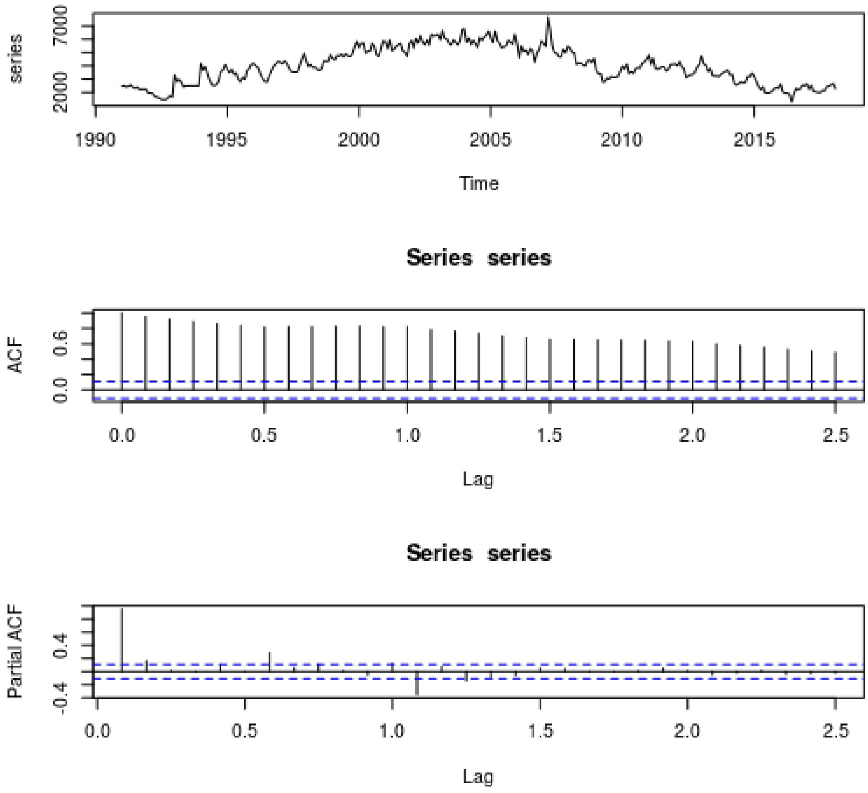


Figure 13: TC093s.txt plots

As it is apparent from Fig. 13 there is a global trend in the series which makes it non-stationary. Performing the stationarity tests confirmed this fact as follows:

```
Dickey-Fuller = -2.3124, Truncation lag parameter = 5, p-value = 0.4448
```

Listing 10: Phillips-Perron Unit Root Test

```
Dickey-Fuller = -1.3798, Lag order = 6, p-value = 0.8381
alternative hypothesis: stationary
```

Listing 11: Augmented Dickey-Fuller Test

As it can be seen in the above outputs of the PP and ADF tests the p-value is not smaller than the critical value 0.01 and hence the series is proved to be non-stationary. Since both of the series in the ARIMAX need to be stationary the next step is to determine which difference is going to make this series stationary. Similar to the previous experiments by looking at the behavior of the ACF with different differencing, it turned out that the seasonal difference is the best one and makes the series stationary. Result of seasonal differencing with `period=12` is shown in Fig. 14.

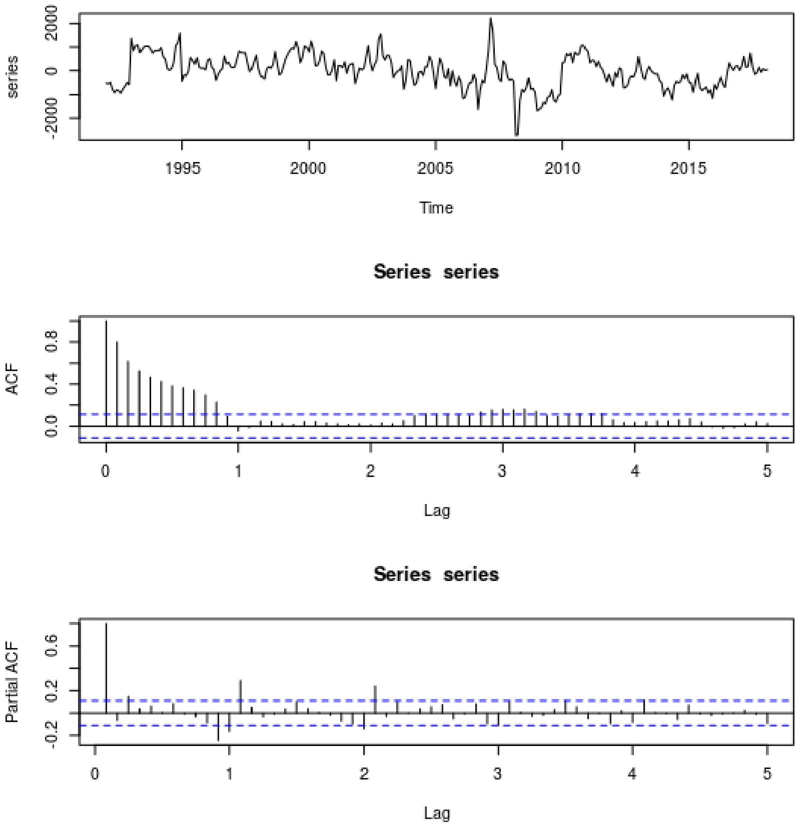


Figure 14: Seasonal Difference of `period=12` for the TC1422s

Now in order to determine the appropriate number of lags to be used from the regressor we need to check the cross correlation of two time series. According to the Fig. 15 we can see that most probably lag-1 will be appropriate. In order to make sure we will employ the pre-whitening approach.

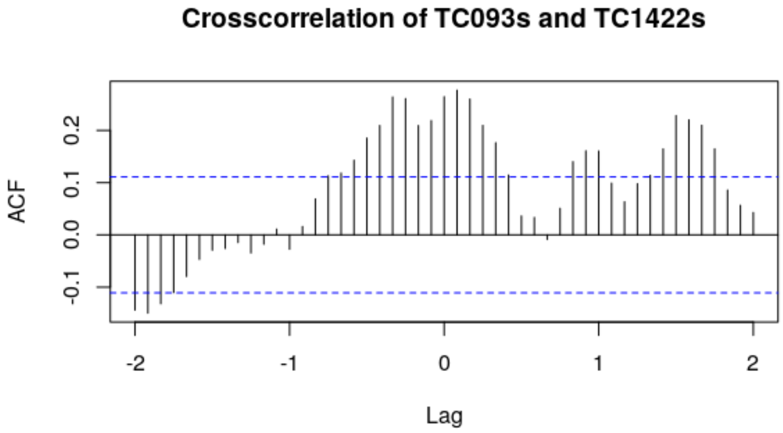


Figure 15: Cross-correlation of two time series TC093s and TC1422s

Fig. 16 shows the cross correlation of the residuals obtained from fitting a same model to both. The chosen model was ARIMA(3,0,1)x(1,1,0) since it was the best model on X. Based on the ccf plot we can see that lag-1 is appropriate.

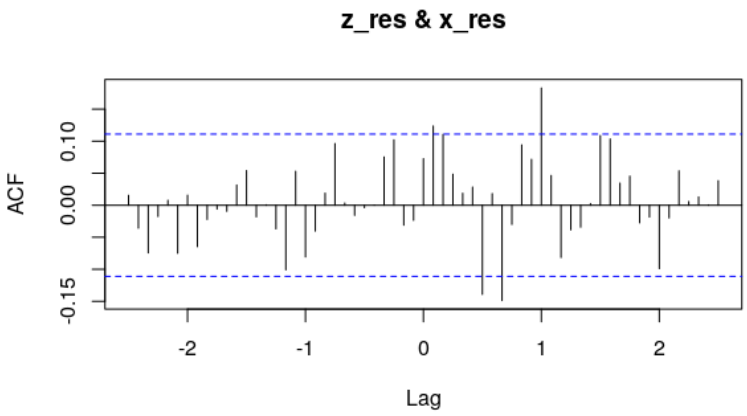


Figure 16: Cross-correlation of residuals of the fitted model to X and Z (pre-whitening method)

After using the lag-1 and fitting an ARIMA model to the Z the following residuals obtained:

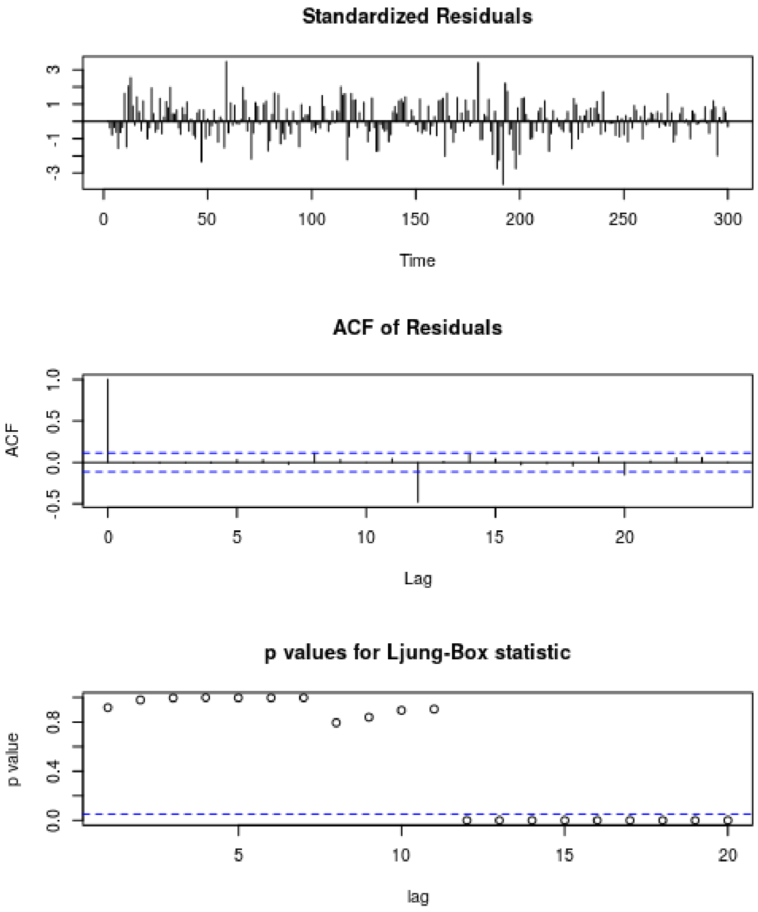


Figure 17: Residual obtained from fitting ARIMA to the Z with X as a regressor

Based on residual and its ACF and PACF depicted in Fig. 17 `ARIMAX(3,0,1)x(1,1,0)` used which its prediction results shown in Fig. 18. According to the Ljung-Box statistic this model is not reliable for lags after 12. Also apparent from the ACF lag 12 which is significant over the limit.

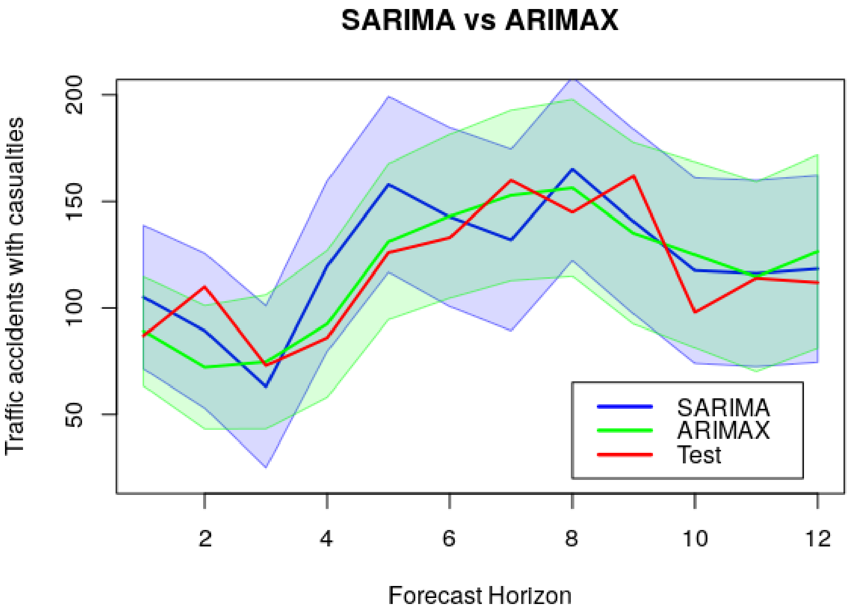


Figure 18: Result of the forecast with the `SARIMA(3,0,0)x(2,1,1)` and `ARIMAX(3,0,1)x(1,1,0)`

2.3 Conclusion of analyzing time series TC093s and TC1422s

I could not obtain better results with ARIMAX. Error of the predictions for the ARIMAX can also be found in Table 3. As can be observed from the forecast in Fig. 18 and the errors table, ARIMAX was not able to achieve better results than SARIMA. Which means a better regressor is needed to achieve better results. To this end, the ARIMAX mathematical formulation is also as follows:

$$\begin{aligned} Z_t &= \beta X_t + \varepsilon_t && ARIMAXgeneralform \\ \Phi(B^{12})\phi(B)\nabla_{12}^{1}\varepsilon_t &= \Theta(B^{12})\theta(B)A_t && \varepsilon_t \text{ fitted to SARIMA model better} \end{aligned}$$

```
arima(x = z1, order = c(3, 0, 1), seasonal = c(1, 1, 0), xreg = x1)

Coefficients:
        ar1 ar2 ar3 ma1 sar1 x1
     0.4540 0.1442 0.1626 -1.0000 -0.0572 0.0084
s.e. 0.3752 0.2111 0.0894 0.0105 0.3798 0.0034

sigma^2 estimated as 691.5: log likelihood = -1398.96, aic = 2811.91
```

Listing 12: Fitted ARIMAX Model output

$$(1-\Phi_1 B^{12})(1-\phi_1 B-\phi_2 B^2-\phi_3 B^3)(\varepsilon_t-\varepsilon_{t-12}) = (1+\theta_1 B)A_t$$

$$\begin{aligned} (\varepsilon_t-\varepsilon_{t-12}) - \phi_1(\varepsilon_{t-1}-\varepsilon_{t-13}) - \phi_2(\varepsilon_{t-2}-\varepsilon_{t-14}) - \phi_3(\varepsilon_{t-3}-\varepsilon_{t-15}) - \Phi_1(\varepsilon_{t-12}-\varepsilon_{t-24}) \\ +\Phi_1\phi_1(\varepsilon_{t-13}-\varepsilon_{t-25}) + \Phi_1\phi_2(\varepsilon_{t-14}-\varepsilon_{t-26}) + \Phi_1\phi_3(\varepsilon_{t-15}-\varepsilon_{t-27}) = A_t + \theta_1 A_t \end{aligned}$$

$$\begin{aligned}\varepsilon_t = {} & \phi_1(\varepsilon_{t-1} - \varepsilon_{t-13}) + \phi_2(\varepsilon_{t-2} - \varepsilon_{t-14}) + \phi_3(\varepsilon_{t-3} - \varepsilon_{t-15}) - \Phi_1(\varepsilon_{t-12} - \varepsilon_{t-24}) - \\ & \Phi_1\phi_1(\varepsilon_{t-13} - \varepsilon_{t-25}) - \Phi_1\phi_2(\varepsilon_{t-14} - \varepsilon_{t-26}) - \Phi_1\phi_3(\varepsilon_{t-15} - \varepsilon_{t-27}) + A_t + \theta_1 A_t + \varepsilon_{t-12}\end{aligned}$$

$$\begin{aligned}Z_t = {} & \beta X_t + \phi_1(\varepsilon_{t-1} - \varepsilon_{t-13}) + \phi_2(\varepsilon_{t-2} - \varepsilon_{t-14}) + \phi_3(\varepsilon_{t-3} - \varepsilon_{t-15}) - \Phi_1(\varepsilon_{t-12} - \varepsilon_{t-24}) - \\ & \Phi_1\phi_1(\varepsilon_{t-13} - \varepsilon_{t-25}) - \Phi_1\phi_2(\varepsilon_{t-14} - \varepsilon_{t-26}) - \Phi_1\phi_3(\varepsilon_{t-15} - \varepsilon_{t-27}) + A_t + \theta_1 A_t + \varepsilon_{t-12}\end{aligned}$$

$$\begin{aligned}Z_t = {} & 0.0084X_t + 0.4540(\varepsilon_{t-1} - \varepsilon_{t-13}) + 0.1442(\varepsilon_{t-2} - \varepsilon_{t-14}) + 0.1626(\varepsilon_{t-3} - \varepsilon_{t-15}) - \\ & (-0.0572)(\varepsilon_{t-12} - \varepsilon_{t-24}) - (-0.0572 \times 0.4540)(\varepsilon_{t-13} - \varepsilon_{t-25}) - \\ & (-0.0572 \times 0.1442)(\varepsilon_{t-14} - \varepsilon_{t-26}) - (-0.0572 \times 0.1626)(\varepsilon_{t-15} - \varepsilon_{t-27}) + \\ & A_t + (-1.0000)A_t + \varepsilon_{t-12}\end{aligned}$$