Time-Series 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 monthly mean value of surface waters at given sites and locations.

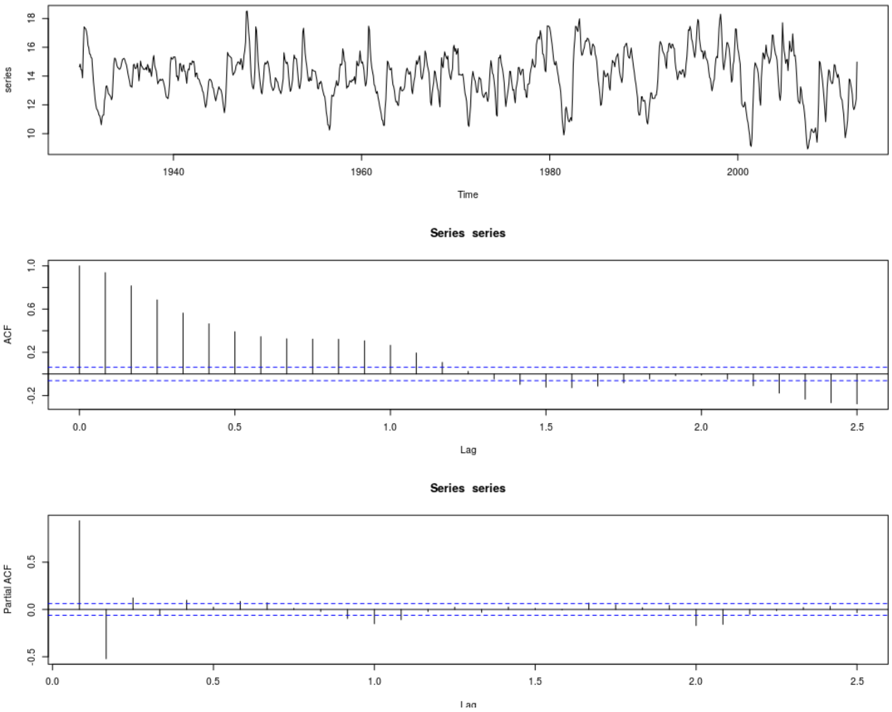
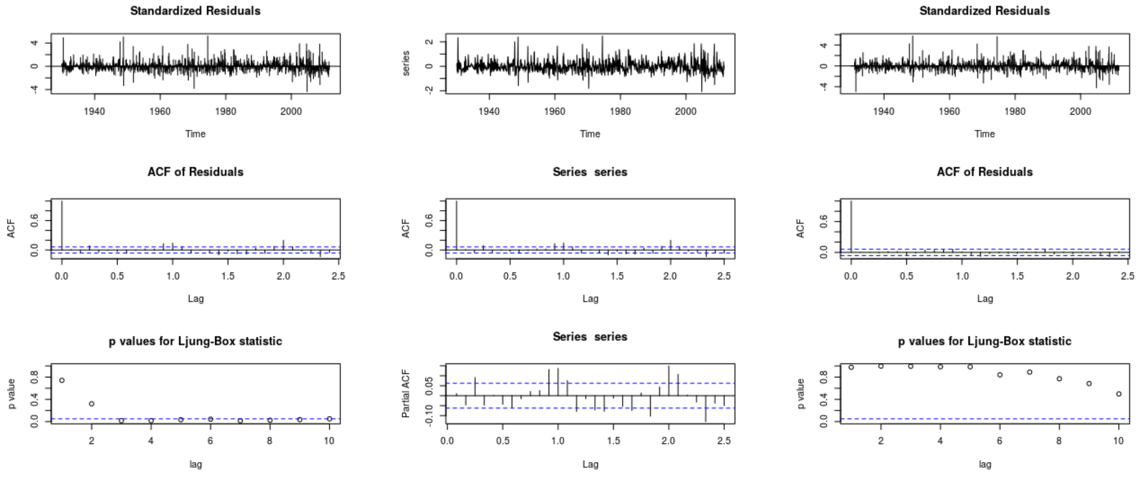


Figure 1: WL45.txt plots

In Fig. 1 we can see that there is no trend in the data. From the PACF we can see that there are most probably 3 AR terms existing.

1.1 (S)ARIMA family

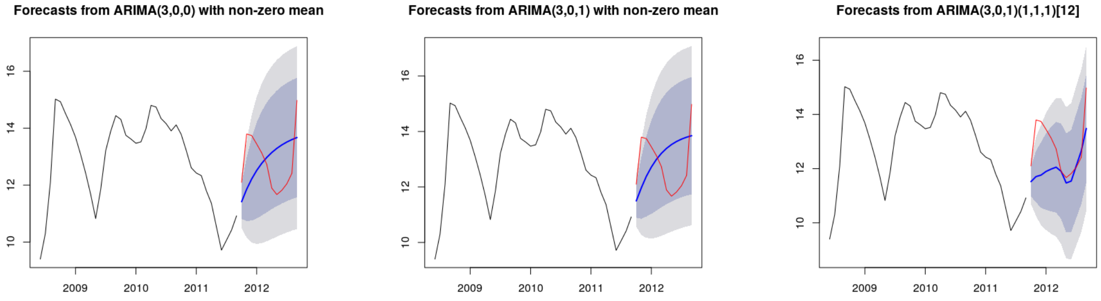
Based on Fig. 1 we start with ARIMA model with AR(3). Based on results obtained from this model (depicted in Fig. 2-(a) we can see that perhaps a MA(1) exist. Study of the ACF and PACF of the residuals from the ARIMA(3,0,1) suggests for a seasonal term. By trial and error ARIMA(3,0,1)x(1,1,1) seems to give the best results. To find the seasonal part (1,1,1) the error measurements that is also reported in Table 1 were considered. Also study of the residuals based on Ljung-Box statistic suggests that the model is appropriate however, it seems that some additional AR/MA terms could also be added.



(a) `ARIMA(3,0,0)` (b) `ARIMA(3,0,1)` (c) `SARIMA(3,0,1)x(1,1,1)`

Figure 2: Diagnosis of the ARIMA/SARIMA fitting on WL45 using `tsdiag`

Point forecast and prediction intervals of these models depicted in Fig. 3 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)`

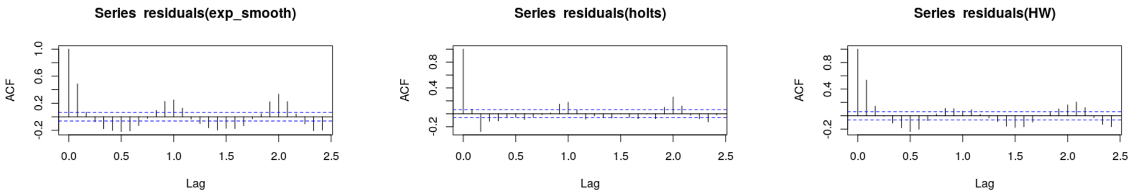
Figure 3: Forecast of the ARIMA/SARIMA fitted models on WL45

	MAE	MSE	MAPE
ARIMA(3,0,0)	1.176005	1.278660	9.229816
ARIMA(3,0,1)	1.207765	1.332862	9.567123
SARIMA(3,0,1)x(1,1,1)	0.857519	1.126478	6.362497

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

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 as well for the comparison. Based on Fig. 4 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. 5.



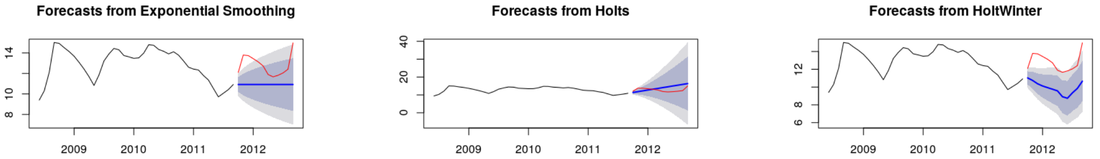
(a) `Exponential Smoothing`

(b) `Holts`

(c) `Holts Winter`

Figure 4: Residuals of the HW-family models studied with auto-correlation function

As depicted in Fig. 5-(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. I tried with different transformation approaches such as log-transform and differencing but the result is more or less the same.



(a) `Exponential Smoothing`

(b) `Holts`

(c) `Holts Winter`

Figure 5: Forecast of the HW-family models

	MAE	MSE	MAPE
Exponential Smoothing	1.894033	2.130382	14.307822
Holt	1.854273	2.163968	14.874172
Holt Winter	2.891893	2.983005	22.401574

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

1.3 Conclusion of analyzing time series WL45

The best model from (S)ARIMA family was SARIMA(3,0,1)x(1,1,1) as it gained the minimum errors among all other. Also from the second part we can conclude that the Holt-winter was the best model. However, SARIMA model performed even better than the holt-winter.

2 Second time series

The second time series that needed to be analyzed was "TC093s" which was the monthly data for the traffic accidents with casualties. As we can see in the Fig. 6 the series is not stationary and has seasonality that is obvious both from the time series plot and the acf plot.

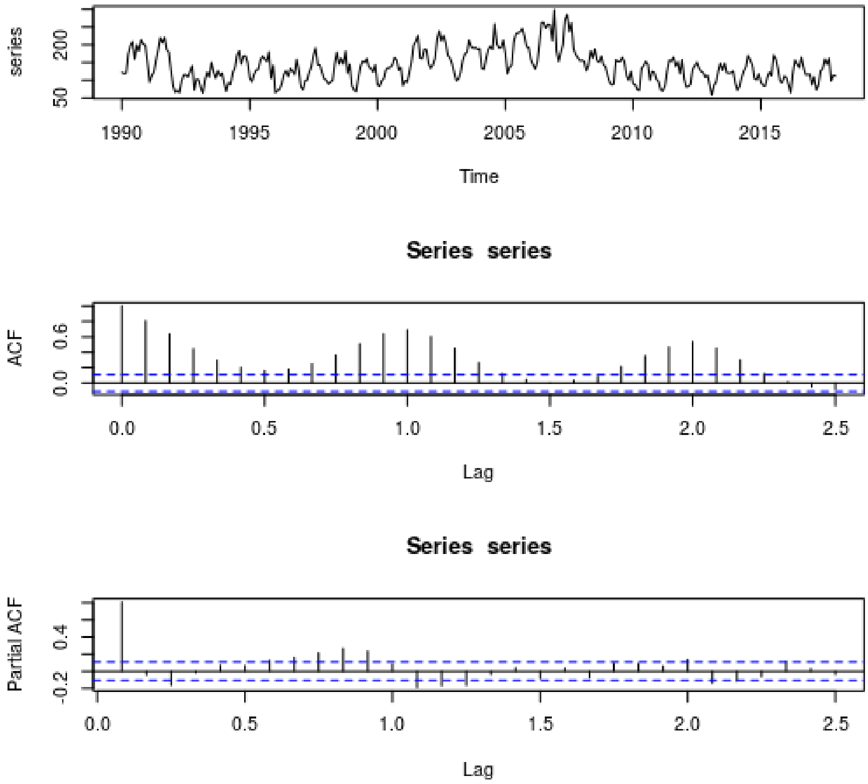
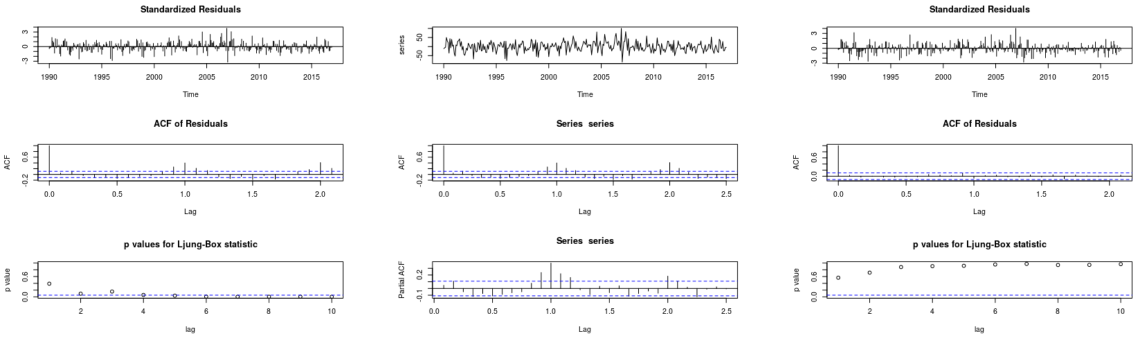


Figure 6: TC093s.txt plots

2.1 (S)ARIMA family

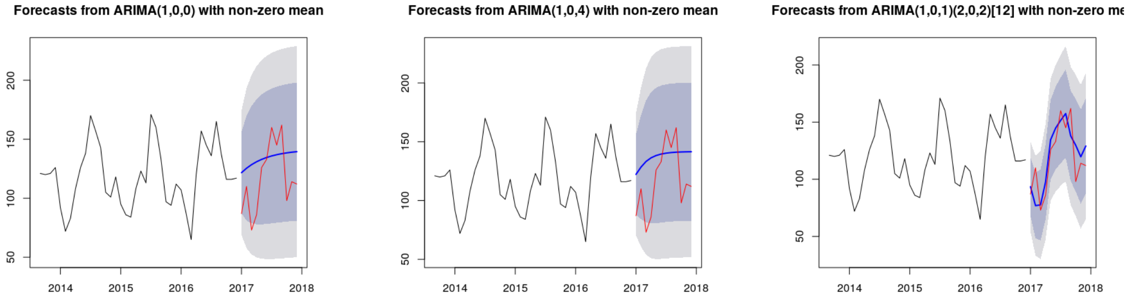
Due to existence of obvious auto-regression term, recognizable from the PACF plot in Fig. 6 I chose to fit an ARIMA(1,0,0) model. By studying the result depicted in Fig. 7-(a) it is also obvious that a MA term exists. I tried MA(4) because of the slightly significant peak on the ACF but it is obvious that there is a sesaonality as we can see there are recurring peaks on the AR(2) and MA(2) and hence a SARIMA(1,0,1)(2,0,2) chose.



(a) `ARIMA(1,0,0)` (b) `ARIMA(1,0,4)` (c) `SARIMA(1,0,1)x(2,0,2)`

Figure 7: Diagnosis of the ARIMA/SARIMA fitting on TC093s using `tsdiag`

Results obtained from each model depicted in Fig. 8 for the forecasting and the errors in Table 3.



(a) `ARIMA(1,0,0)` (b) `ARIMA(1,0,4)` (c) `SARIMA(1,0,1)x(2,0,2)`

Figure 8: Forecast of the ARIMA/SARIMA fitted models on TC093s

	MAE	MSE	MAPE
ARIMA(1,0,0)	25.68667	30.04117	25.99806
ARIMA(1,0,4)	27.33929	31.98421	27.97023
SARIMA(1,0,4)x(2,0,2)	14.63566	17.41611	12.86195
ARIMAX(3,1,3)x(2,1,2)	24.20700	29.40927	22.48880

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

According to the forecasting plots and the table we can see that the best model was SARIMA(1,0,1)x(2,0,2).

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 Dec 2018. In order to align the starting and ending dates of both I dropped the first 12 months of the Z and 12 last months of the X. Now what we would like to do in the ARIMAX model is to come up with a model that predicts the response variable based on different predictors.

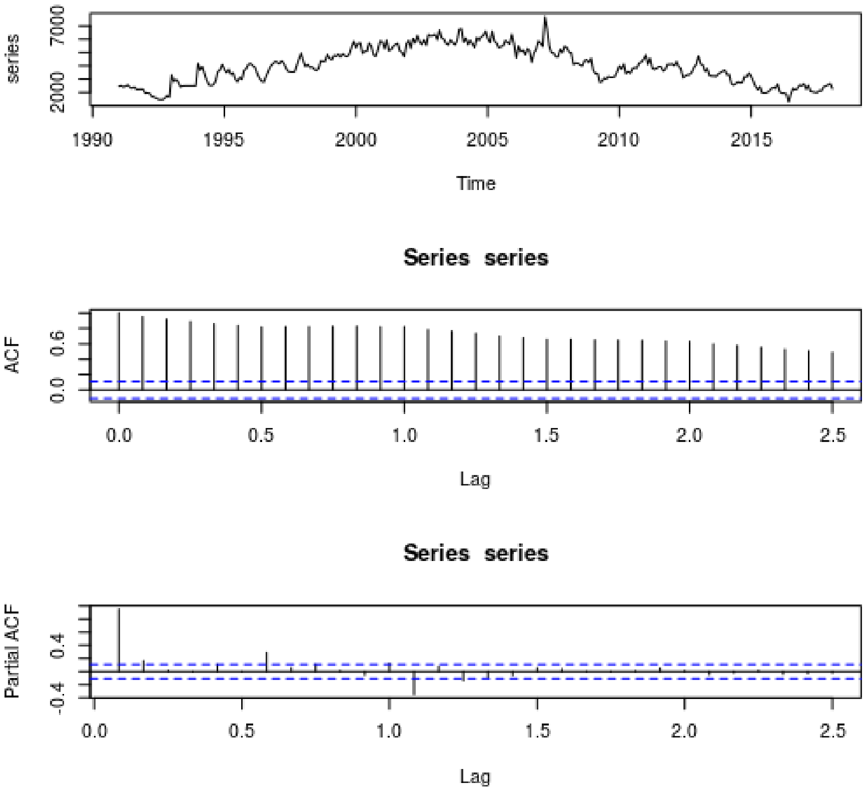


Figure 9: TC093s.txt plots

First think that we would like to do is to see the cross correlation of two time series. According to the Fig. 10 we can see that there is a continuous correlation as the lag increases. However, from this plot it is not obvious which lag we should use.

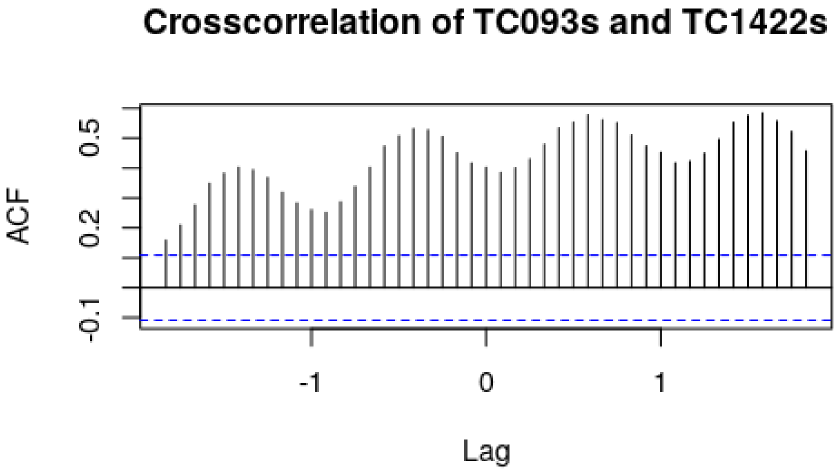
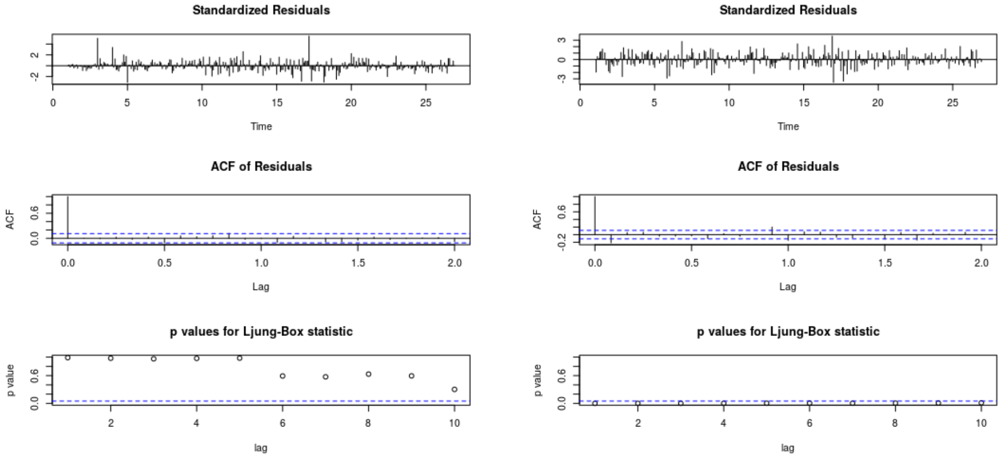


Figure 10: Cross-correlation of two time series TC093s and TC1422s

For the next step I chose to fit the best model on TC1422s (X) and then try to fit the same model on TC093s (Z) with the coefficients from the model fitted on the previous time series. Result of the this fitting shown in Fig. 11. The best model on X was ARIMA(2,1,1)x(2,0,0) with log transformation of the input. The same model used with coefficients obtained from fitting the model to X, in order to fit this time on Z. The Fig. 11-(a) shows the first and (b) the second model fitting results. As it can be seen in Fig. 11-(b) the model is not suitable.



(a) `fitting on` X_t

(b) `fitting on` Z_t

Figure 11: Result of fitting a model on the predictor ts and the target ts with coefficients of the previous model

Since the previous experiment was not successful, I tried to find the appropriate number of lags that should be used from X in order to make prediction for Z using the prewhitening approach. Fig. 12 shows the cross correlation of the residuals obtained form fitting a same model to both. The model was again ARIMA(2,1,1)x(2,0,0) since it was the best model on X. Based on the ccf plot we can see that one lag is enough, but again it seems that the X is not explaining Z very much and may not help in improving the results.

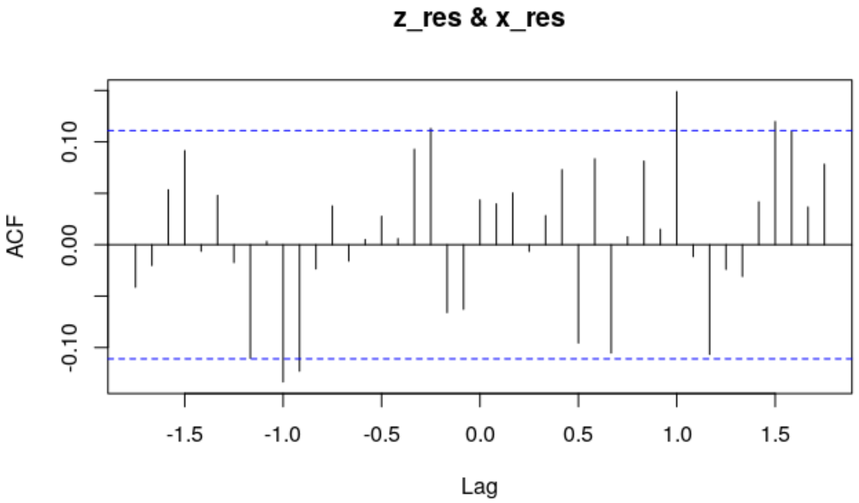


Figure 12: Cross-correlation of residuals of the fitten model to X and Z (prewhitening method)

After using the lag1 and fitting an ARIMA model to the Z the following residuals obtained:

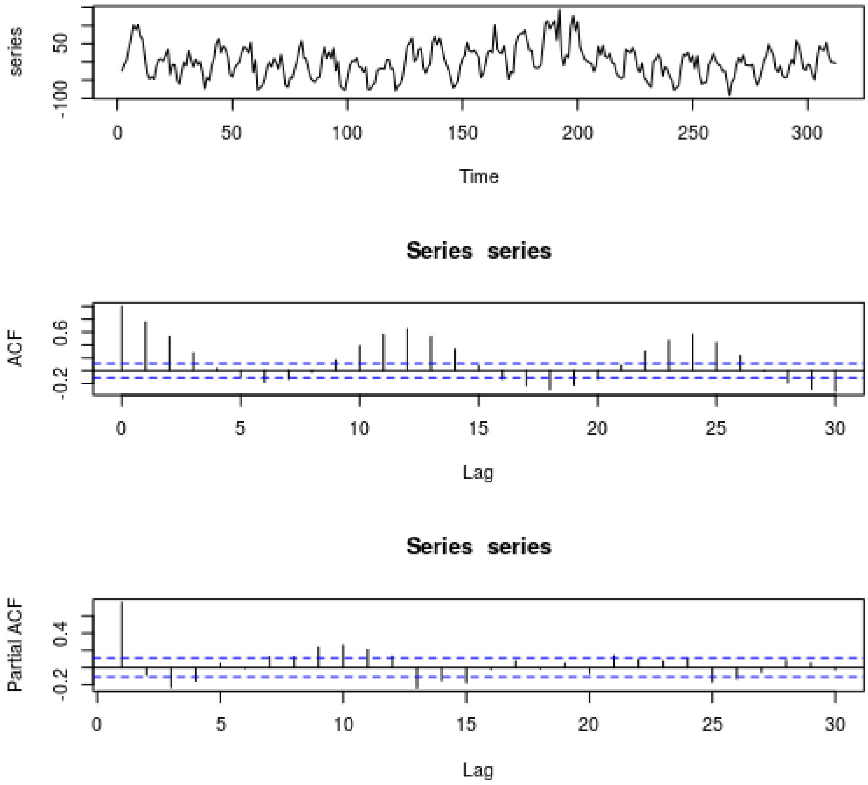


Figure 13: Residual obtained from fitting ARIMA to the Z with X as external regressor

Based on residual and its ACF and PACF depicted in Fig. 13 we used SARIMA(3,1,3)x(2,1,2) which its prediction results shown in Fig. 14.

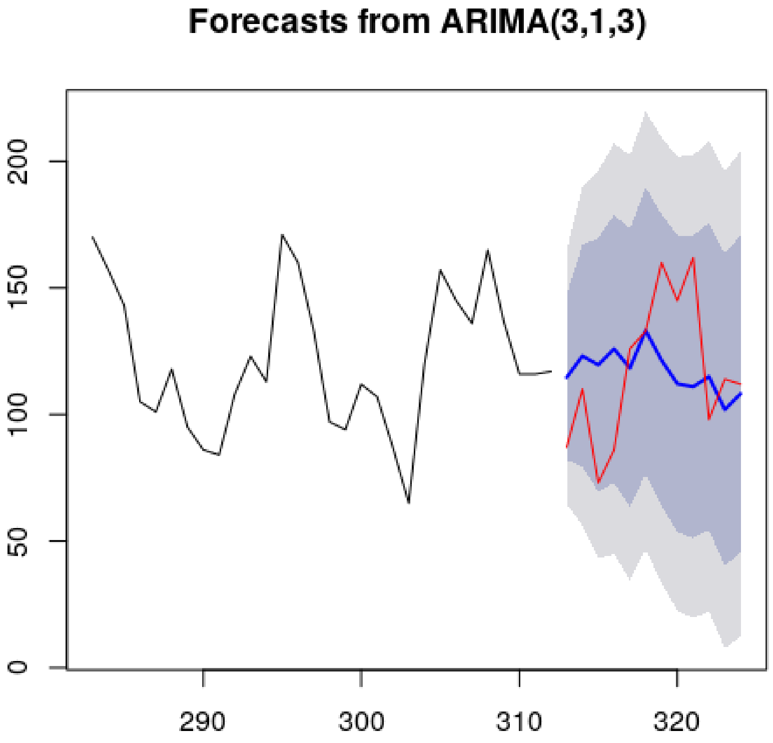


Figure 14: Result of the forecast with the ARIMAX(3,1,3)x(2,1,2)

2.3 Conclusion of analyzing time series TC093s and TC1422s

I could not obtain better results with ARIMAX. It was mainly because the external time series that I intended to use was not explaining the target time series. This was shown with the cross correlation and prewhitening method. Also, the TC1422s was the only time series that had its date matching with TC093s. In general the best model for TC093s that obtained a reasonable result was SARIMA(1,0,4)x(2,0,2) as reported in Table 3.