Analyzation and Model Selection of EQcount Dataset

Yao Yao

University of Toronto

STA457

Professor Tharshanna Nadarajah

April 16, 2021

Abstract

In this report, I analyzed the time-series dataset EQcount(Series of annual counts of major earthquakes (magnitude 7 and above) in the world between 1900 and 2006), proposed two possible models, and made predictions for the next 10 years. I first proposed two models: model(0,1,1) and model(2,1,0) through plotting the autocorrelation function(acf) and the partial autocorrelation function(pacf). By performing a diagnostics check and comparing the estimated parameters, we determined that model(0,1,1) is more suitable with our selection criteria. Then I forecasted the data into future ten years ahead through this model, and found that the counts of major earthquakes will actually not have too many fluctuations in the next ten years. Also, by finding the first three predominant periods through the periodogram, we could conclude that we cannot distinguish the spectrum well and establish significance of three peaks.

Introduction

People have been suffering from natural disasters for a long time, one of the disasters is major earthquakes. The definition of major earthquakes is the earthquakes with magnitude 7 and above, and major earthquakes would cause a huge damage to people. Thus, if we could forecast the future situations by using statistical analysis, we could use this as a basis to help us prevent the damage of earthquakes to some extent or determine whether our actions have an impact on the environment. Therefore, I chose the dataset EQcount under the library astsa. EQcount is a series of annual counts of major earthquakes in the world between 1900 and 2006. Based on this dataset, our main goal is to propose a suitable model and use that model to forecast the count of major earthquakes in the next ten years. We would also identify the first three predominant periods based on this dataset.

Statistical Method

In order to propose a model, we could use the autocorrelation function(acf) and the partial autocorrelation function(pacf) of a stationary series to determine the order of the model. First, we have to determine whether the original data is stationary. If the series is not stationary, we could take some transformations such as differencing and log transformation to make the series stationary. Then we plot the acf and pacf, which would provide the order of MA model and AR model. In this case, the order of MA model is 1 and the order of AR model is 2, and order of difference is 1. Based on this, we proposed two SARIMA models, ARIMA(0,1,1) and ARIMA(2,1,0). In order to find a better model, we could perform the diagnostics. We would check whether the standard residuals are showing no trend, the points in acf of residuals are within the blue lines, the model is under normality and p-values are all greater than 0.05. If all the proposed models passed, then we compare the AIC and BIC, and choose the smaller one. Also, we could check the significance of the parameters, and choose a model with all parameters significant.

Then, we use the appropriate model to forecast the next 10 years and calculate 95% prediction intervals for each of the ten years. Last, we draw the periodogram to identify the first three predominant periods and obtain the 95% confidence intervals for the identified periods

Results

By plotting the original EQcount series, we found that this series is not stationary. It has a clear upward trend before 1950 and a downward trend after 1950. (Figure 1). Thus it may require a difference. Figure 2 shows the series after differencing. After making a difference, we found that the amplitude is inconsistent, thus this may require a log transformation. Then after making

a difference and log transformation, the plot is shown as Figure 3. We could see that the series fluctuates up and down around a line, which demonstrates that it is stationary.

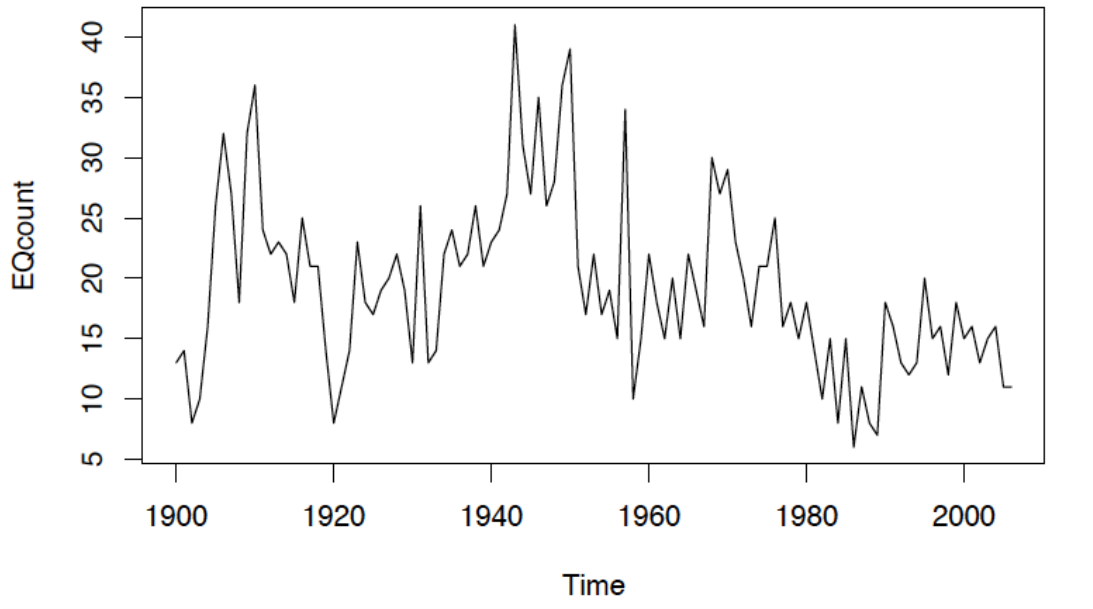

Figure 1. The original EQcount series

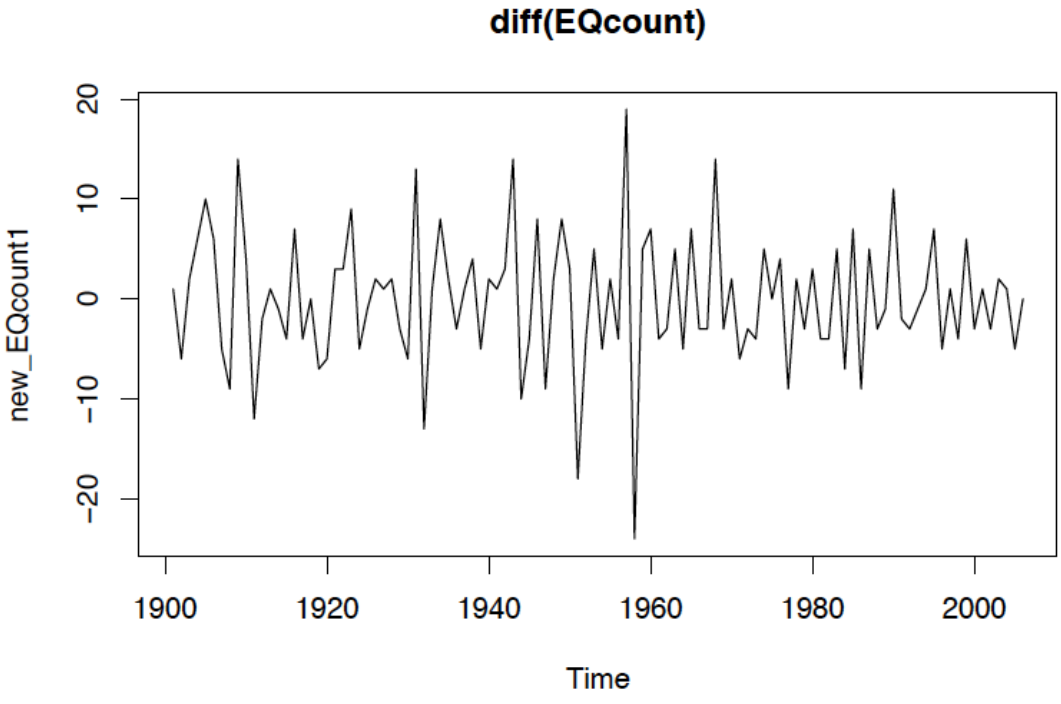

Figure 2. The EQcount series after taking difference

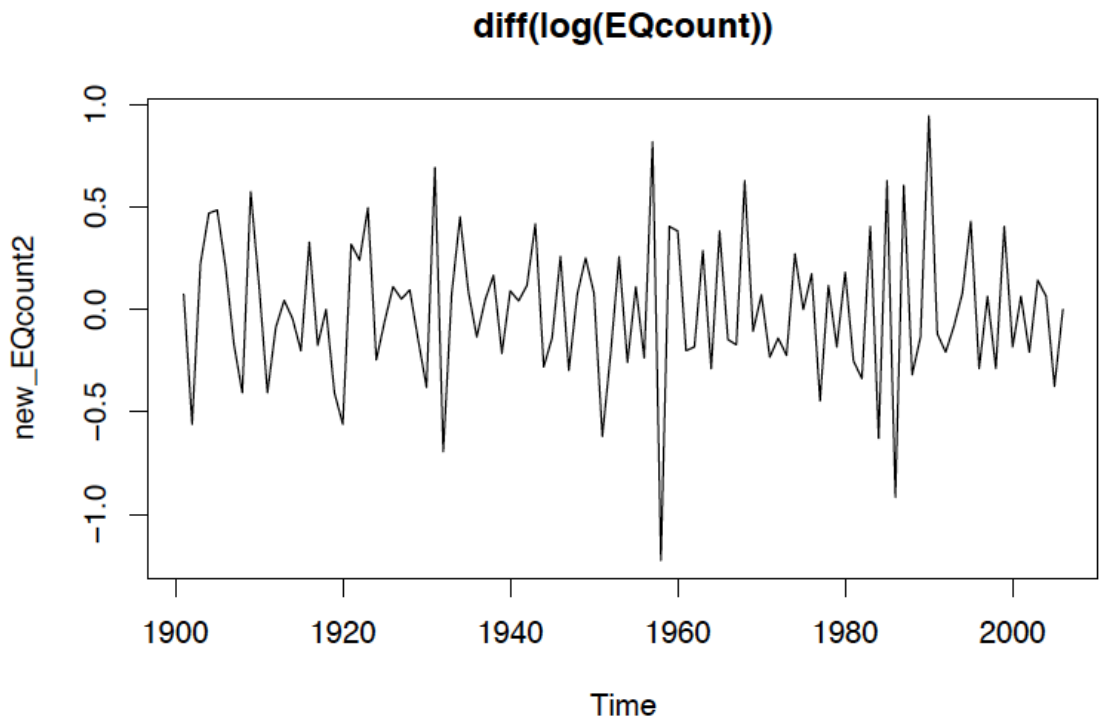

Figure 3. The EQcount series after taking difference and log transformation

Then we plot the acf and pacf of the new data(Figure 4 and Figure 5). The cutoff of acf in Figure 4 is 1, this demonstrates a MA(1) model. The cutoff of pacf in Figure 5 is 2, this demonstrates a AR(2) model. Based on this, we could propose two ARIMA models. ARIMA(0,1,1) and ARIMA(2,1,0) according to the original data.

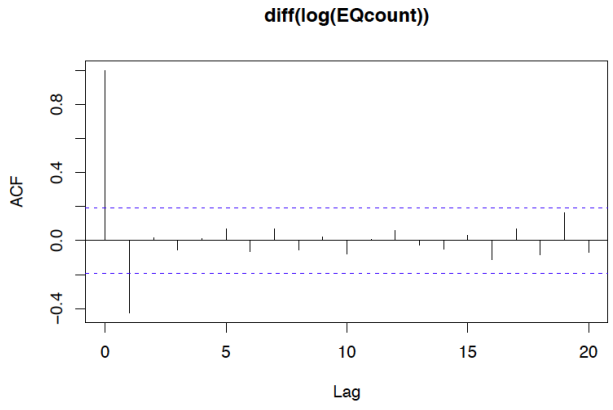

Figure 4. ACF of new EQcount series

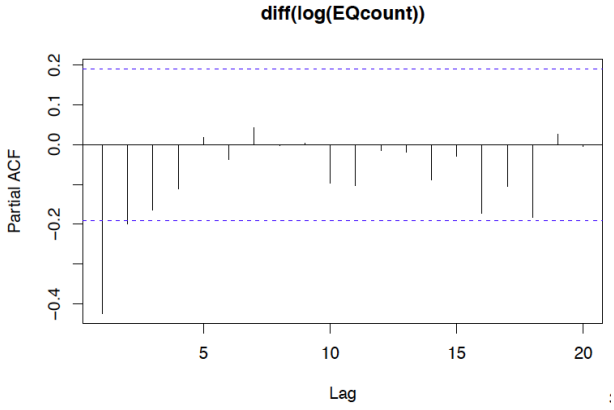

Figure 5. PACF of new EQcount series

Figure6 and Figure7 are the diagnostic checks for both models.

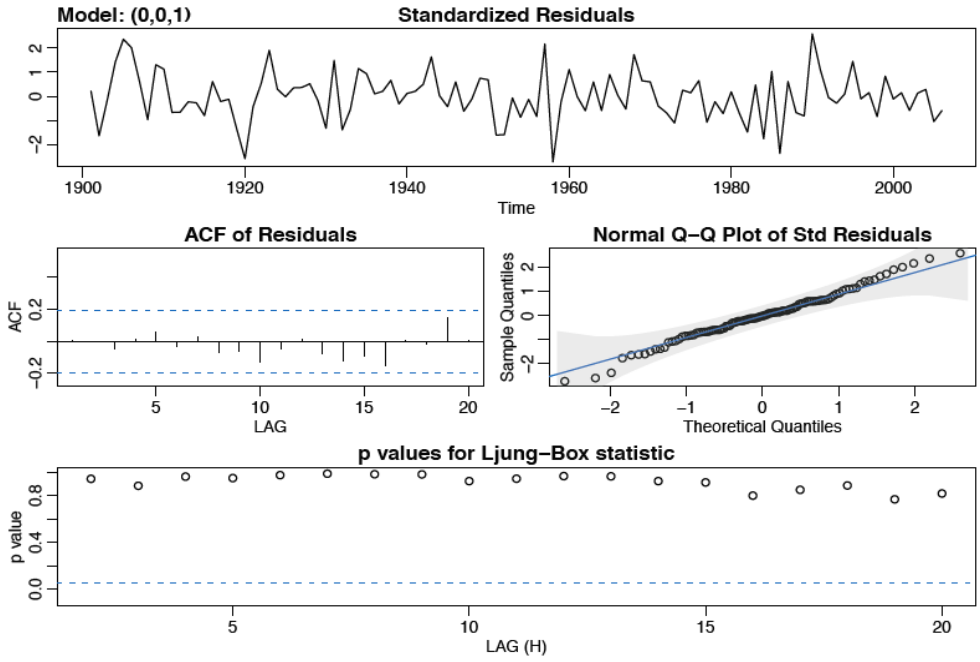

Figure 6. diagnostics of model(0,0,1)

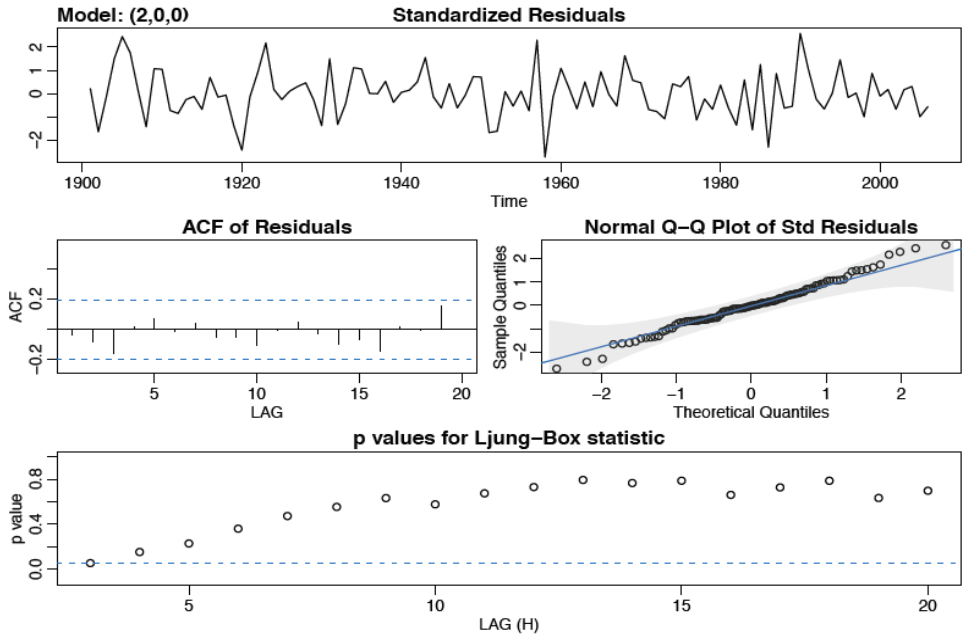

Figure 7. diagnostics of model(2,0,0)

- The standardized residuals for both plots are showing no trend, this both of them are white noise.
- The ACF of Residuals: All points are within the blue lines, thus the residuals look white.
- The QQ plot of the residuals: The model is under normality but with the exception of the possible outliers.
- p-values for Ljung-Box: the p-value are all greater than 0.05, which shows there is no evidence against the null hypothesis
- Referring to AIC and BIC, the AIC of ARIMA(0,0,1) is 0.5511948, while ARIMA(2,0,0) is 0.6028634. Thus AIC refers to ARIMA(0,0,1). The BIC of ARIMA(0,0,1) is 0.6265752, while ARIMA(2,0,0) is 0.7033705. Thus BIC refers to ARIMA(0,0,1).
- Through significant tests, the p-value of parameter in ARIMA(0,0,1) is 8.049e-12 smaller than 0.05, which provides strong evidence against the null hypothesis. The p-value of parameters in ARIMA(2,0,0) are 9.423e-08 and 0.03504 smaller than 0.05, which also provides strong evidence against the null hypothesis.
- Estimated parameter for ARIMA(0,0,1) is theta1:-0.567115. Thus W_(t-1) and Wt would have impact on Xt

  Estimated parameters for ARIMA(2,0,0) are phi1:-0.5069018, phi2:-0.20243. Thus X_(t-1) and X_(t-2) would have impact on Xt

Overall, we propose ARIMA(0,1,1) model which is based on the original data, (ARIMA(0,0,1) is based on the new data) and use that for the forecasting. Figure 8 shows the forecasting for the next 10 years. Table 9 shows the prediction interval for each of the 10 years.

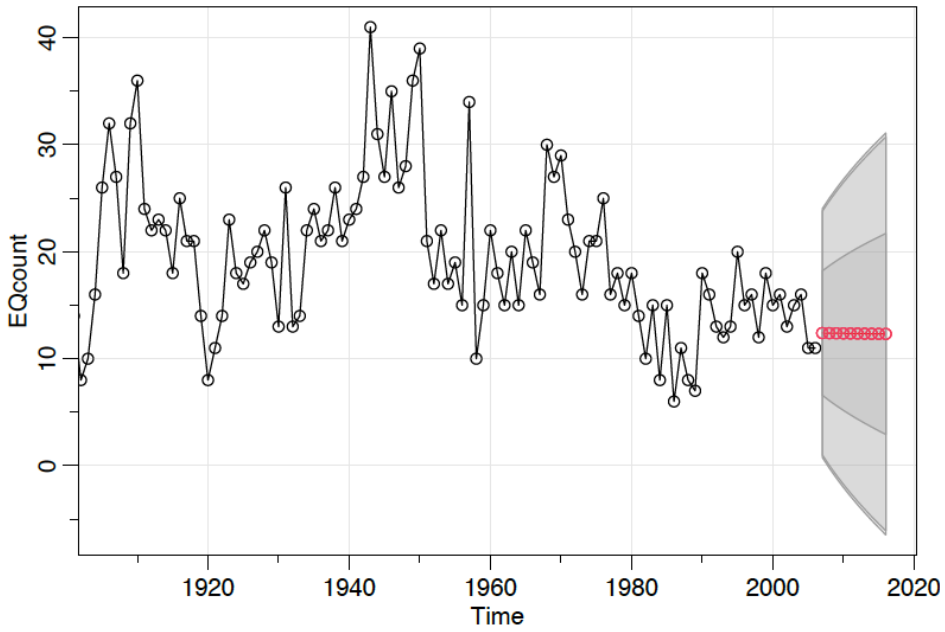

Figure 8. Forecasting for next 10 years

	L	U
2007	1.01867361	23.77600
2008	0.02990962	24.74671
2009	-0.88669402	25.64525
2010	-1.74506331	26.48557
2011	-2.55514919	27.27759
2012	-3.32435511	28.02874
2013	-4.05836616	28.74469
2014	-4.76166046	29.42993
2015	-5.43784007	30.08805
2016	-6.08985367	30.72201

Table 9. prediction interval for each of the 10 years.

From the forecasting and the prediction interval, we could find that the counts of major earthquakes will actually not have too much variation in the next ten years. Also the prediction interval is gradually wider, this shows the fluctuations are gradually greater.

Then, we plot the periodogram(Figure 10) to find the dominant spectrum and confidence interval(Table 11)

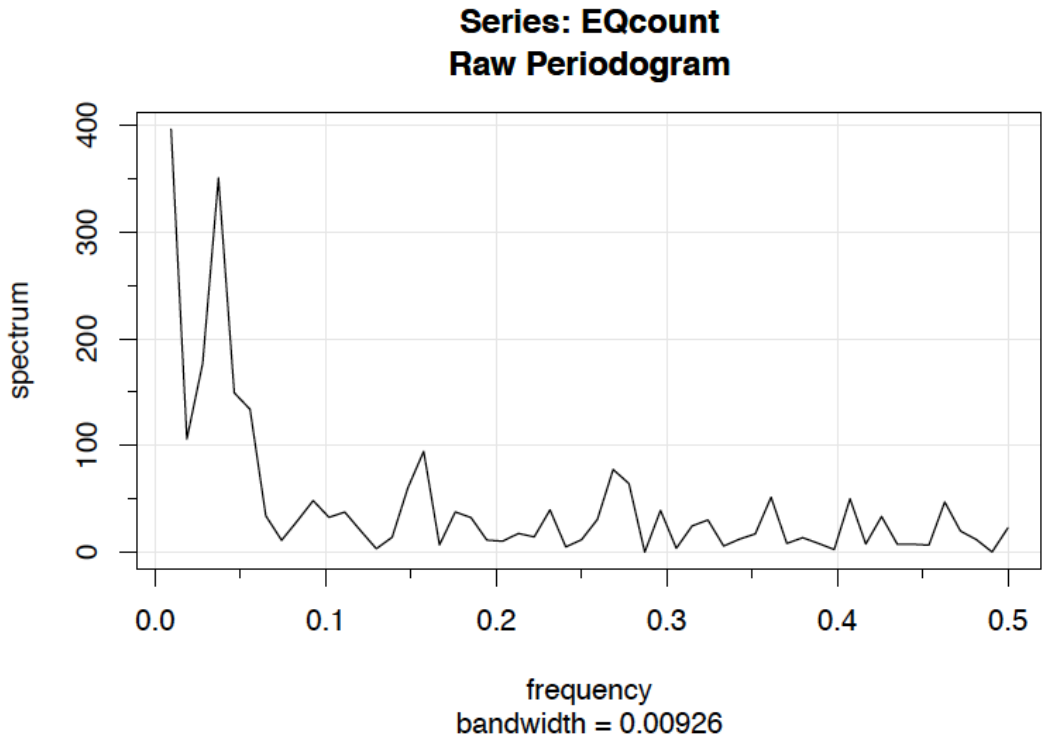

Figure 10. Periodogram of EQcount series

	spectrum	upper	lower
first	396.2778	15652.137	107.42498
second	350.7439	13853.644	95.08142
third	176.8623	6985.688	47.94472

Table 11. dominant spectrum and confidence interval

We cannot establish significance of all the peaks, since the peaks lie in the confidence interval of all three periods. Example, the periodogram ordinate is 396.2778, which lies in the confidence intervals of the second and third peaks, thus we cannot establish significance of the first peak. Similarly, we cannot establish significance of the second peak and third peak. Also, we could conclude that we cannot distinguish the spectrum well because of the overlapping confidence intervals.

Discussion

There seems to be some seasonal trend. So multiplicative seasonal ARIMA models may have been a better fit. Also, some outliers are detected at the tails of the Q-Q plot limiting the model prediction. Besides, the sample size of EQcount is only 107. A sample with small size may not be so representative.