STA457 Final Project

Predicting annual counts of major earthquakes using ARIMA model based on EQcount data set

Ruike Xu 1006562550

15/04/2022

Contents

Abstract

There is no doubt that earthquakes are dangerous and unpreventable natural hazards that could happen to us. There are no scientists that can successfully predict the accurate occurence of earthquakes. This analysis aims to study the annual counts of major earthquakes from 1900 to 2006 globally and predict the possible counts of major earthquakes in the future. We focus on ARIMA model construction based on the EQcount data set, data forecast for the next ten years and its corresponding 95% predictive intervals, and spectral analysis of the first three predominant periods. We proposed two possible models (ARIMA(2, 1, 1) and ARIMA(0, 1, 1)) for our differencing earthquakes data set while ARIMA(0, 1, 1) is diagnosed as the best model of representation. The analysis indicates that the annual counts of major earthquakes would stay roughly constant for the next ten years.

Keywords: Earthquakes, ARIMA, Data forecast, Confidence interval, Spectral analysis, data transformation, Natural hazards

Introduction

Earthquake is one of the most hazardous disasters that could occur on Earth. Earthquake destruction starts with the violently shaking of the Earth, which could cause landslides and fluidify the Earth's surface. Countless devastating blows could happen to the local people where the earthquake occurs. (CEA, 2020) Earthquakes can be caused by a wide range of causes, which includes mining on the surface and underground, extraction of fluids and gas from the subsurface, etc. (USGS, 2022) However, scientists can only provide a probability that a major earthquake would occur in a specific area within a certain number of years. Our study aims to predict the annual counts of major earthquakes (defined as

magnitude 7 and above) from 2007 to 2016 based on the EQcount dataset. The data is recorded annually and there are 107 data points in total. (CRAN) By analyzing EQcount we could discover more about the potential occurrence of major earthquakes and take strategic precautions before the occurrence. In this study, we aim to predict the next ten years of annual counts of major earthquakes using an ARIMA model and analyze the EQcount data set using spectral analysis.

Statistical Methods

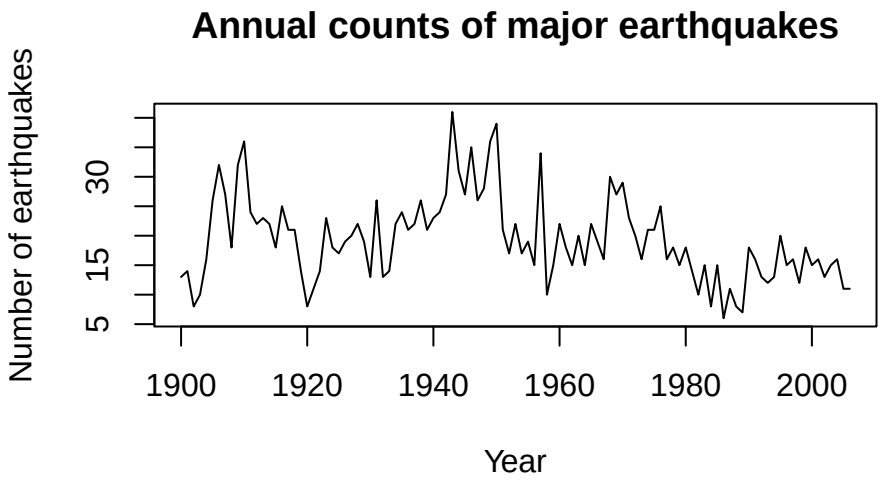

Figure 1: Time series plot for annual counts of major earthquakes from 1900 to 2006

The above plots are based on 107 observations of the global annual counts of major earthquakes from 1900 to 2006. As shown in Figure 1, the annual earthquake counts time series isn't a stationary process due to the non-constant mean. We can also observe there is a general increasing trend from 1900 to 1940 and declined afterward. In Figure 2, the sample ACF doesn't decay to 0 quickly as the time lag h increases, which indicates that differencing is needed as a data transformation method.

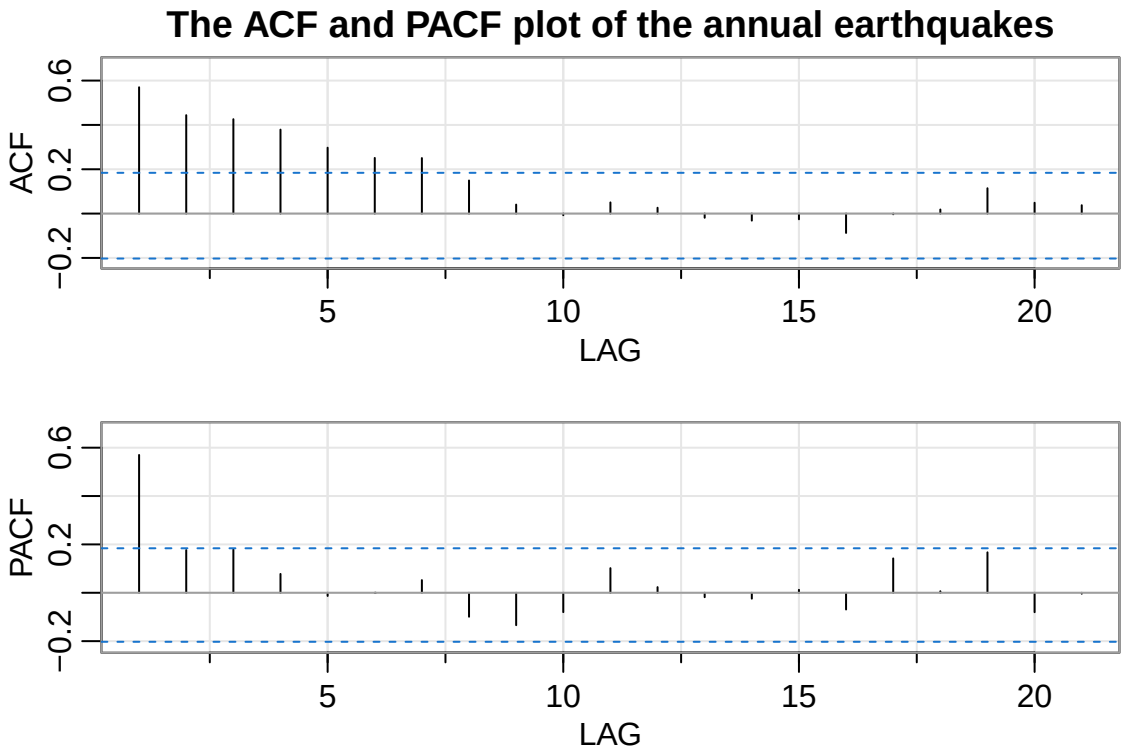

Figure 2: ACF and PACF plots for global annual earthquakes from 1900 to 2006

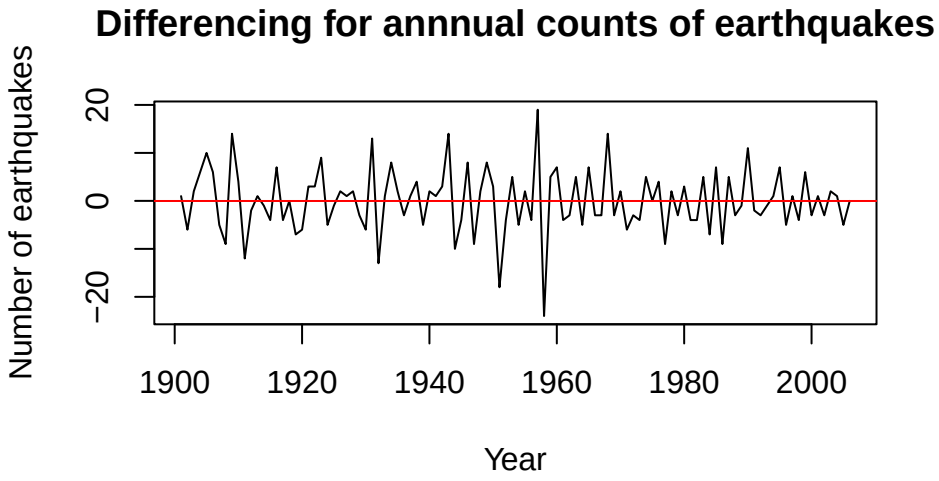

Figure 3: Time series plot for differencing of annual counts of major earthquakes from 1900 to 2006

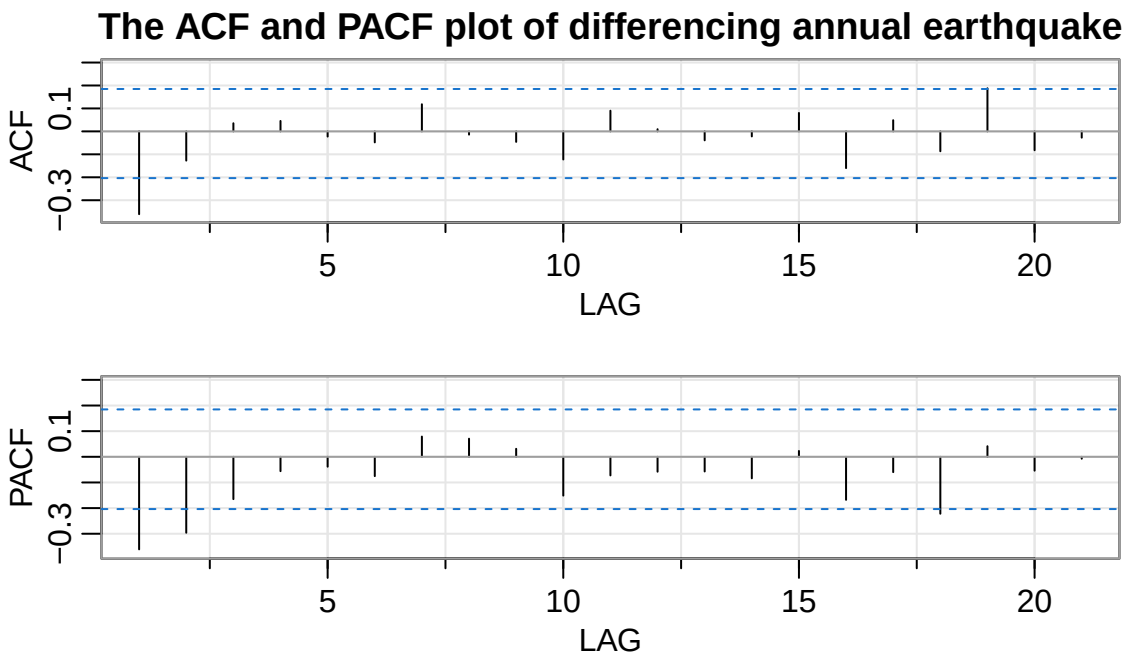

Figure 4: ACF and PACF plots for differencing of global annual earthquakes from 1900 to 2006

We can observe from Figure 3 that the mean of annual counts of earthquakes is roughly constant and close to 0 after we conduct the differencing transformation. (The red line is the mean of the time series) The variance of the transformed time series only depends on time lag. Compared to the previous ACF and PACF plots, the ACF and PACF decay faster and approach 0 after we transform the data using differencing. Thus we can conclude that the differencing of the annual counts of major earthquakes is a stationary process.

Model selection

We can determine the parameters of the ARIMA model that is suitable for the annual counts of major earthquakes by examining the properties of ACF and PACF plots. We can observe from Figure 4 that ACF tails off after lag 1 and PACF tails off after lag

2. There is no significant seasonal pattern of the annual counts of major earthquakes after differencing, which means that we don't need further seasonal adjustment for our data. Based on previous data diagnostics, we propose two models: ARIMA(2, 1, 1) and ARIMA(0, 1, 1)

Result

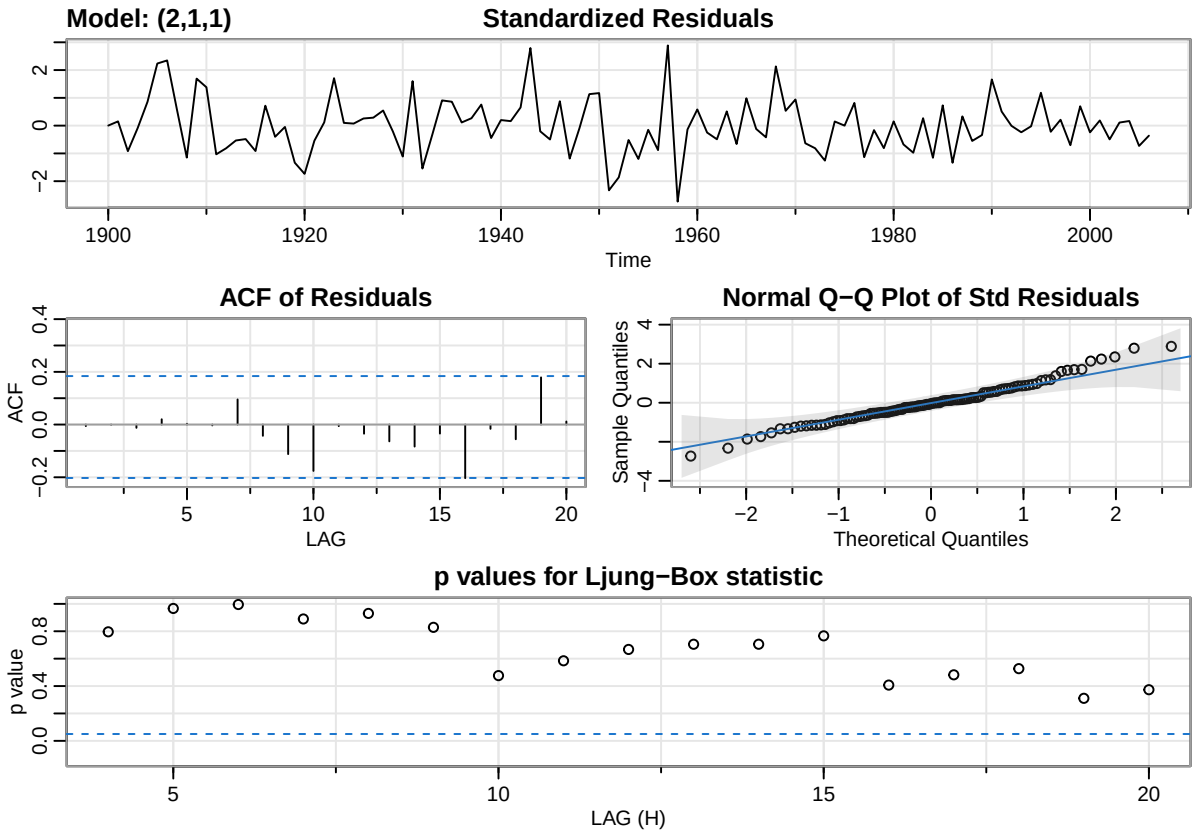

Figure 5: Model diagnostics of ARIMA(2, 1, 1) model

The first ARIMA model we fitted given the annual counts of major earthquakes is ARIMA(2, 1, 1), which is expressed as

$$\bigtriangledown(x_t) = -0.0263_{0.2273} \bigtriangledown (x_{t-1}) - 0.1409_{0.1431} \bigtriangledown (x_{t-2}) - 0.4983_{0.2186} w_{t-1} + w_t$$

	Estimate	SE	t.value	p.value
ar1	-0.0263	0.2273	-0.1157	0.9081
ar2	-0.1409	0.1431	-0.9842	0.3273
ma1	-0.4983	0.2186	-2.2793	0.0247

Table 1: Summary statistics for ARIMA(2, 1, 1) model

We can observe that the coefficients of ar1 and ar2 p_values are not significant at the 0.05 significance level. Hence we don't reject the null hypothesis that these two coefficients are equal to 0. Based on this finding, we could analyze the second model we proposed.

From the above model diagnostics plots, we could observe that the standard residual plot shows no obvious pattern and there are no outliers that are above 3 standard deviation level for the residual. The residual ACF plot has no spikes exceeding the blue line, which indicates that the residual of model 1 moves like a white noise process. The Normal QQ plot indicates no significant departure from the diagonal line (with only a few outliers), which means that the normal model assumption is satisfied. All data points in the Ljung-Box statistics plot are above the 0.05 level, so we fail to reject the null hypothesis that the residuals are independent.

The second ARIMA model we fitted given the annual counts of major earthquakes is ARIMA(0, 1, 1), which is expressed as

$$\bigtriangledown(x_t) = -0.5762_{0.0838} w_{t-1} + w_t$$

The P_value of the coefficient of the model is 0.0000, which is less than the 0.05 level of significance. This implies that we can reject the null hypothesis that the coefficients of the model equal 0.

We perform the model diagnostics for model 2 similar to the previous model. From the below model diagnostics plots, we could observe that the standard residual plot demon-

strates no obvious pattern and there are no outliers that are above 3 standard deviation level for the residual. The residual ACF plot has no spikes exceeding the blue line. The Normal QQ plot indicates no significant are approximately aligned with the diagonal line, with a few more outliers departing compared to the previous model. The Normal QQ plot shows that the overall normal model assumption holds for model 2. All data points in the Ljung-Box statistics plot are above the 0.05 level, so we fail to reject the null hypothesis that the residuals are independent.

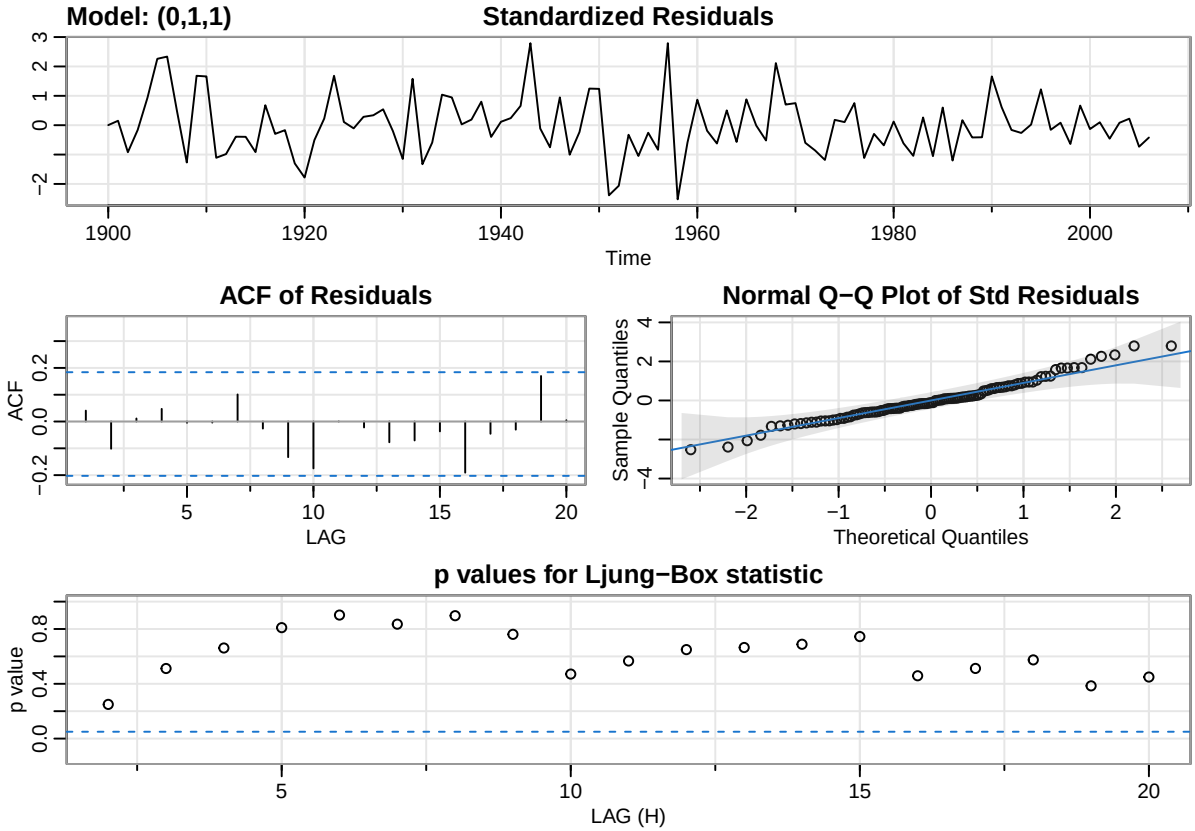

Figure 6: Model diagnostics of ARIMA(0, 1, 1) model

	Estimate	SE	t.value	p.value
ma1	-0.5762	0.0838	-6.8718	0.0000

Table 2: Summary statistics for ARIMA(0, 1, 1) model

Model selection

We would employ AIC, BIC, and AICc as model selection criteria to further examine which model is the most appropriate for our time series. AIC (Akaike's Information Criteria) is a metric that penalizes the inclusion of additional variables to a model, which adds penalties and increases the error when including additional terms. A lower level of AIC indicates a better model. BIC (Bayesian Information Criteria) is a variant of AIC with a stronger penalty for including additional variables in the model. AICc is another version of AIC that is more suitable for small sample size.

Information Criteria	Model 1: ARIMA(2, 1, 1)	Model 2: ARIMA(0, 1, 1)
AIC	6.4402	6.4159
BIC	6.4439	6.4170
AICc	6.5658	6.4913

Table 3: Model selection criteria for two proposed models

As we can observe from the above summary table for information criteria of two proposed models, model 2 performs better than model 1 in every perspective of assessment. On the other hand, since model 1 has two model parameter coefficients that are not significant under the 0.05 level of significance, we choose ARIMA(0, 1, 1) as the best model for further data forecast and spectral analysis.

Data forecast for next ten years and 95% prediction intervals

In the above Figure 8, we can observe from the predictive time series plot that the prediction for the next ten years shows a general constant trend. There is also an increasing uncertainty of predictive intervals as a prediction as the prediction time is away from now. The above Table 3 shows the future ten years' annual counts of major earthquakes and the corresponding 95% predictive intervals from 2007 to 2016. The 95% predictive

Year	Prediction	Lower bound	Upper bound
1	12.40	1.02	23.78
2	12.39	0.03	24.75
3	12.38	-0.89	25.65
4	12.37	-1.75	26.49
5	12.36	-2.56	27.28
6	12.35	-3.32	28.03
7	12.34	-4.06	28.74
8	12.33	-4.76	29.43
9	12.33	-5.44	30.09
10	12.32	-6.09	30.72

Table 4: Prediction for the next ten years with 95% predictive intervals

interval indicates that real prediction is 95% likely to be included in the range of the upper and lower bounds. To be more specific, for the annual counts of major earthquakes in 2007, the real prediction is 95% likely to be included in the range of 1.02 to 23.8.

Spectral analysis of first three predominant periods and corresponding confidence intervals

	Predominant.Freq	Period	Spectrum	Lower_CI	Upper_CI
1	0.0093	108	396.2778	107.4250	15652.137
2	0.0370	27	350.7439	95.0814	13853.644
3	0.0278	36	176.8623	47.9447	6985.688

Table 5: Summary table for first three predominant periods and corresponding confidence intervals

The first three predominant periods are 108, 27, and 36 whereas their corresponding frequencies are 0.0093, 0.0370, and 0.0278. We could observe that the three predominant period's spectrum falls into all three 95% confidence intervals individually, which means that we can't establish the significance of all three peaks.

Discussion

In our analysis of annual counts of major earthquakes, we fit an ARIMA(0, 1, 1) model for the EQcount data set and employ data forecast for the next ten years of major earthquakes counts. We also conduct spectral analysis of the EQcount data set and analyze the predominant periods. In our fitted model, the differencing of this year to the previous year's annual count is negatively associated with the white noise. We can observe that for the future ten years the annual counts of major earthquakes behave a constant trend. Based on our research, we could further take strategic precautions prior to the occurrence of earthquakes.

There are certainly some limitations to our research. First, the data set contains only 107 observations, which could result in a biased sample and an inaccurate estimation of our research question. Second, the annual counts of major earthquakes are recorded based on the occurrence of the global, which means that the prediction of the earthquakes wouldn't be regionally accurate based on our analysis. It's also worth mentioning that a record of counts of seasonal occurrence would further improve the accuracy of prediction.

In the future study, we could perform a similar analysis with a more recent, monthly-recorded regional data to specify the potential occurrence of major earthquakes in a certain region. Studying the monthly-record data could potentially uncover a seasonal pattern of the major earthquakes.

Reference

Kassambara. (2018, March 11). Regression Model Accuracy Metrics: R-square, *AIC, BIC, CP and more*. STHDA. Retrieved from http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/

USGS. (2022). Can we cause earthquakes? Is there any way to prevent earthquakes? | U.S. Geological Survey. (n.d.). Retrieved from https://www.usgs.gov/faqs/can-we-cause-earthquakes-there-any-way-prevent-earthquakes

CEA - earthquake damage, Danger & Destruction. Earthquake Damage - How Earthquakes Cause Danger & Destruction | CEA. (2020, August 10). Retrieved from https://www.earthquakeauthority.com/Blog/2020/How-Earthquakes-Cause-Damage-Destruction

Comprehensive R Archive Network (CRAN). (n.d.). *Package astsa*. CRAN. Retrieved from https://cran.r-project.org/web/packages/astsa/

Sarima: Fit arima models. RDocumentation. (n.d.). Retrieved from https://www.rdocumentation.org/packages/astsa/versions/1.14/topics/sarima

Xie, Y. (2017, December 3). The R package tinytex - Yihui Xie: Helper Functions to Manage TinyTeX, and Compile LaTeX Documents. Retrieved from https://yihui.org/tinytex/r/