MAT5314 Graduate Project Team 2

Teng Li(7373086)
Shiya Gao(300381032)
Chuhan Yue(300376046)
Yang Lyu(8701121)

Introduction

Earthquakes are usually caused when the underground rock suddenly breaks and there is rapid motion along a fault. They can be categorized into natural earthquakes and artificial earthquakes. Natural earthquakes can be further classified into tectonic earthquakes and volcanic earthquakes. Tectonic earthquakes are caused by the rupture of rocks deep in the ground and the rapid release of energy accumulated over a long period of time. Volcanic earthquakes are caused by volcanic eruptions. In addition, artificial earthquakes are caused by human activities that alter the stresses and strains on Earth's crust. In our project we will only focus on the natural earthquake due to its probabilistic nature of occurrence.

Earthquakes are difficult to predict for three main reasons. Firstly, plate boundaries are prone to earthquakes, but the way and speed of plate movement is difficult to measure and predict accurately. Secondly, the low frequency of large earthquakes does not provide enough data for earthquake modeling. Thirdly, it is difficult to enter the inner crust of the earth to observe the data, so most of the detection of earthquakes is done by collecting some vibrations and signals on the surface to analyze and predict them, which leads to a decrease in the accuracy of earthquake prediction.

Frequent historical earthquakes give us a huge data set that can be used to study earthquake causes, correlations, space-time, and further earthquake prediction. Current researches have provided several statistical models to analyze the earthquake processes. Trigger model, a special case of the Neyman-Scott clustering model, can be used to estimate aftershocks after a major earthquake (Ogata 1988). Epidemic-Type Aftershocks Sequence (ETAS) model is also widely used to forecast earthquake occurrences (ROSS and KOLEV 2022). In addition, time series analysis can be done to explore cycles related to earthquake frequency and it is also effective in predicting large earthquakes (Amei, Fu, and Ho 2012). There are also several techniques from machine learning that provide us alternative ways other than conventional statistical models to analyze the earthquake data. For example, deep learning can be used to predict seismic events, including intensity and location (Nicolis, Plaza, and Salas 2021). Clustering model can identify regions with high-frequency earthquakes to upgrade building structures to reduce damage from impending earthquakes.

We focus on a data set of earthquake occurrence in Canada provided by the Government of Canada. The data set is located at https://open.canada.ca/data/en/dataset/4cedd37e-0023-41fe-8eff-bea45385e469. In this project we want to analyze the data to give a detailed review of this natural disaster in the country and make some inference about its characteristics. In particular, we want to check if there's any location-wise pattern in earthquake occurrence and how frequent it can happen. Then we want to analyze if the occurrence of earthquake is also correlated to some other variables such as time. Finally we want to apply some probabilistic models that have been extensively studies by other researchers to the data in order to assess the likelihood of future occurrence in high-risk area.

Method

First, we will demonstrate some aspects of the earthquake occurrence by visualizing the seismic data. Map charts and other related plots will be used to perform such initial analysis. Then we will use time series analysis to study whether there exist temporal trends, seasonal patterns, and certain frequency in seismic events. Third, we will use geographic characteristics such as longitude, latitude and magnitude to identify potential groups of clusters. We may use different techniques of cluster analysis to achieve this goal. This will allow us to compare the predicted clusters with the actual location of earthquakes. Lastly, we will try

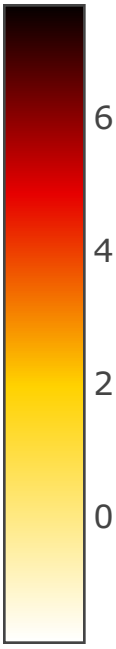
to estimate the probability of earthquakes in high-risk regions by using the variables from the data and the Epidemic-Type Aftershocks Sequence (ETAS) model.

Exploratory Analysis

```
data<-read.csv("eqarchive-en.csv")
head(data)
```

```
##                       date latitude longitude depth magnitude magnitude.type
## 1 1985-01-01T11:01:00+0000   47.000   -66.600   5.0       1.9             MN
## 2 1985-01-01T12:34:37+0000   78.800  -102.330  18.0       3.2             MN
## 3 1985-01-02T00:13:06+0000   48.880  -122.994  22.5       1.2             ML
## 4 1985-01-02T00:56:59+0000   50.022  -121.554  18.0       1.0             ML
## 5 1985-01-02T05:03:03+0000   49.459   -66.752  18.0       1.8             MN
## 6 1985-01-02T06:43:51+0000   50.609  -130.344  10.0       2.8             ML
##                         place  X X.1
## 1 61 km E  from Plaster Rock NB
## 2     26 km E  from Isachsen NU
## 3   21 km SW from White Rock BC
## 4      56 km W  from Merritt BC
## 5     40 km N  from Cap-Chat QC
## 6  202 km W  from Port Hardy BC
```

```
fig <- data%>%plot_ly(
  lat = ~latitude,
  lon = ~longitude,
  marker = list(color = ~magnitude, colorscale = 'Hot', opacity=0.5, size=~magnitude, showscale=TRUE, r
  type = 'scattermapbox', mode="markers",text = ~paste("Magnitude: ", magnitude)) %>%
  layout(
    mapbox = list(
      style = 'open-street-map',
      zoom =2.5,
      center = list(lon = -88, lat = 34)))

fig
```



We have plotted the locations of earthquakes on the map, and it is evident that there is a substantial amount of data. Earthquakes primarily occur in regions of Canada that are in proximity to the United States and Greenland. British Columbia, New Brunswick, and Yukon experience earthquakes more frequently than other provinces. British Columbia and Yukon are located on the Pacific Plate boundary, where the Pacific Plate is subducting beneath the North American Plate. This subduction can lead to significant seismic activity. New Brunswick is near the St. Lawrence rift system, which is associated with fault lines and seismic activity.

The magnitude of earthquakes is usually divided into seven main categories, which, depending on the magnitude, we can describe as follows:

1. “Microearthquake”: Magnitude less than or equal to 1. These are tiny earthquakes that are usually hard to detect.
2. “Minor Earthquake” or “Microseismic”: the magnitude is greater than 1 and less than or equal to 3. This type of earthquake is usually not easy to be perceived, especially when the epicenter is deeper.
3. “Felt Earthquake”: Magnitude greater than 3, less than or equal to 4.5. People can feel this kind of earthquake, but it usually does not cause damage.
4. “Moderate Earthquake”: The magnitude of the earthquake is greater than 4.5, less than or equal to 6. This type of earthquake is potentially destructive, but the extent of damage depends on a number of factors, such as the depth of the epicenter and the distance from the epicenter.
5. “Strong Earthquake”: Magnitude greater than 6, less than or equal to 7. This type of earthquake is more destructive.
6. “Major Earthquake”: The magnitude is greater than 7 and less than or equal to 8. This type of earthquake may cause surface rupture and widespread seismic wave propagation.
7. “Massive Earthquake”: The magnitude is greater than 8 or higher. Such earthquakes usually result in significant crustal displacement and seismic wave propagation, posing a significant threat to the surrounding area.

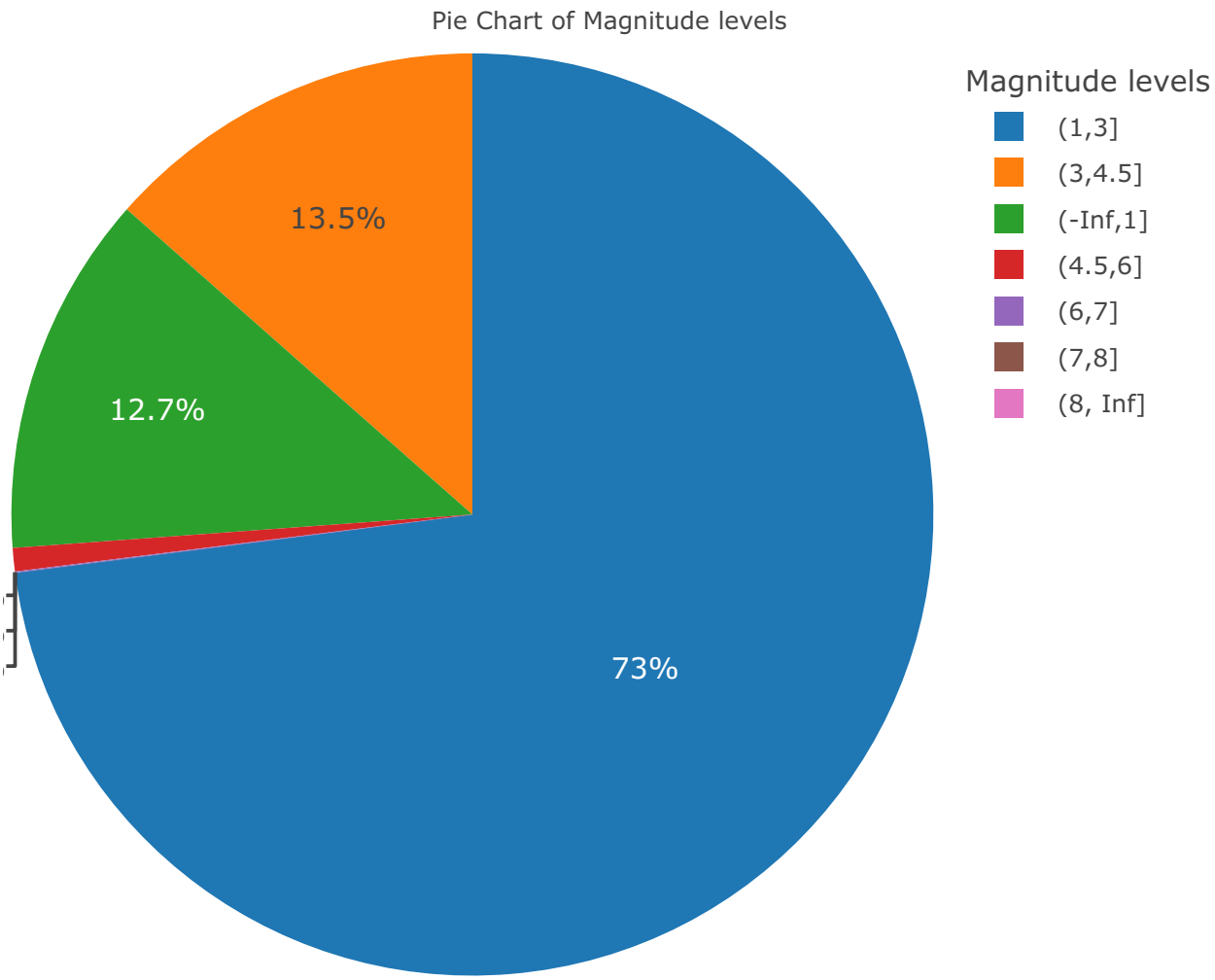
We calculated the number of earthquakes occurring in each of these seven different earthquake magnitude categories separately. We found that between 1985 and 2019, "Moderate Earthquake" occurred 828 times ; "Strong Earthquake" had 36 recorded occurrences; "Major Earthquake" occurred only three times, and "Massive Earthquake" never occurred during this 35-year period. The cumulative number of earthquakes in these categories shows a significant difference compared to the other three categories. This suggests that most earthquakes occurred in these decades did not cause much damage to people, buildings, or the ground.

```
eq_data<-read.csv("eqarchive-en.csv")

#Dividing the magnitude into 7 levels
breaks <- c(-Inf, 1, 3, 4.5, 6, 7, 8, Inf)
categories <- cut(eq_data$magnitude, breaks = breaks)
magnitude_level <- table(categories)
magnitude_level_df <- data.frame(
  Magnitude_Range = as.character(names(magnitude_level)),
  Earthquake_Count = as.numeric(magnitude_level)
)
print(magnitude_level_df)
```

```
##   Magnitude_Range Earthquake_Count
## 1        (-Inf,1]            12842
## 2           (1,3]            73983
## 3         (3,4.5]            13673
## 4         (4.5,6]              828
## 5           (6,7]               36
## 6           (7,8]                3
## 7        (8, Inf]                0
```

In order to compare the relative proportions of the number of earthquakes occurring in each earthquake magnitude category corresponding to the overall number of earthquakes more visually, we plotted the pie chart. As can be seen, 73% earthquake's magnitude $M \in (1, 3]$, which was "Minor Earthquake". This was the highest ratio for these seven earthquake magnitude categories and it accounted for almost three quarters of the total earthquakes happened between 1985 and 2019 year. The proportion of earthquakes whose magnitude $M \in (3, 4.5]$ was a little greater than the earthquakes whose magnitude $M \in (-\infty, 1]$. These two types of earthquakes accounted for 25% of the total and the remaining four categories of large earthquakes accounted for less than 1% of the total. Although these earthquakes were the most destructive, most earthquakes occurred in these decades did not cause much damage to people, buildings, or the ground.

```
#Calculating the percentages for different earthquake magnitude's levels
percentages <- (magnitude_level_df$Earthquake_Count / sum(magnitude_level_df$Earthquake_Count))*100
A <- as.data.frame(percentages)

#Draw the pie chart of the percentage
piechart <- plot_ly(A, labels=~magnitude_level_df$Magnitude_Range, values=~percentages, type="pie", wid
  layout(margin=list(l=0, r=0, b=0, t=20),
         title = list(text="Pie Chart of Magnitude levels", font=list(size=10)),
         xaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
         yaxis = list(showgrid = FALSE, zeroline = FALSE, showticklabels = FALSE),
         legend = list(title = list(text="Magnitude levels"),font=list(size=10))
  )
piechart
```

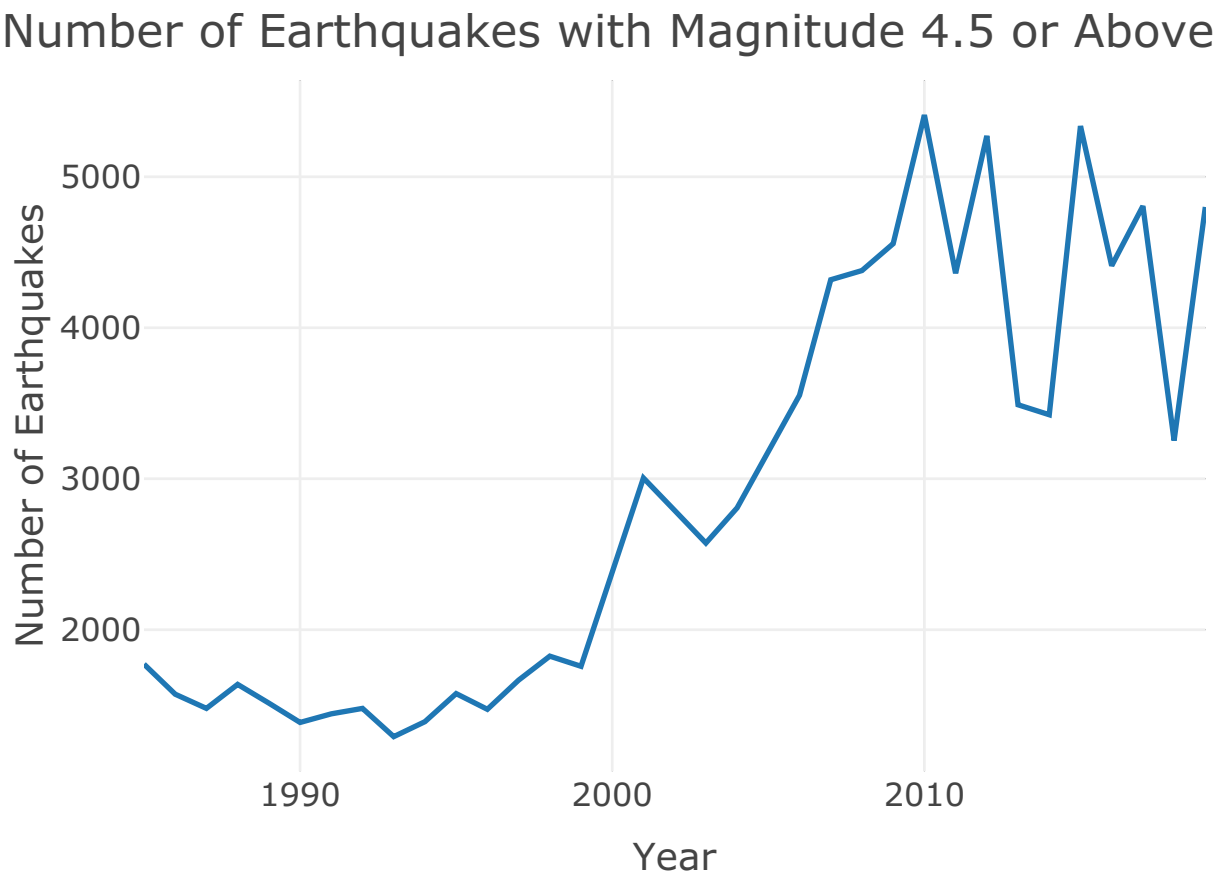


We have separately calculated the number of earthquakes with a magnitude $M \in [4.5, \infty)$ that occurred each year and plotted a time series graph. From 1985 to 1999, there was a slow decrease followed by an increase in the number of earthquakes. However, the annual earthquake count remained around 1500 on average and did not exceed 2000. Starting from the year 2000, we observed a significant upward trend in the occurrence of earthquakes with a magnitude of 4.5 or higher, reaching its peak in 2010 with 5409 events. From 2011 to 2019, although there were noticeable fluctuations in the annual earthquake count, it consistently remained above 3000 and did not drop below 3253, indicating a relatively high frequency.

Based on this, we can infer that from 1985 to 2019, the number of earthquakes with a magnitude of 4.5 or higher increased with the passage of years, particularly during the ten-year period from 2000 to 2010, when the growth rate was particularly pronounced.

```
selected_eq_data <- eq_data[eq_data$magnitude >= 4.5, ]

frequency <- eq_data %>%
  group_by(year = year(date)) %>%
  summarise(count = n())
frequency$year <- as.numeric(frequency$year)

x <- ts(frequency)

plot_ly(data = frequency, x = ~year, y = ~count, type = "scatter", mode = "lines", name = "Number of Ea
  layout(title = "Number of Earthquakes with Magnitude 4.5 or Above",
```

```
        xaxis = list(title = "Year"),
        yaxis = list(title = "Number of Earthquakes")
        )
```

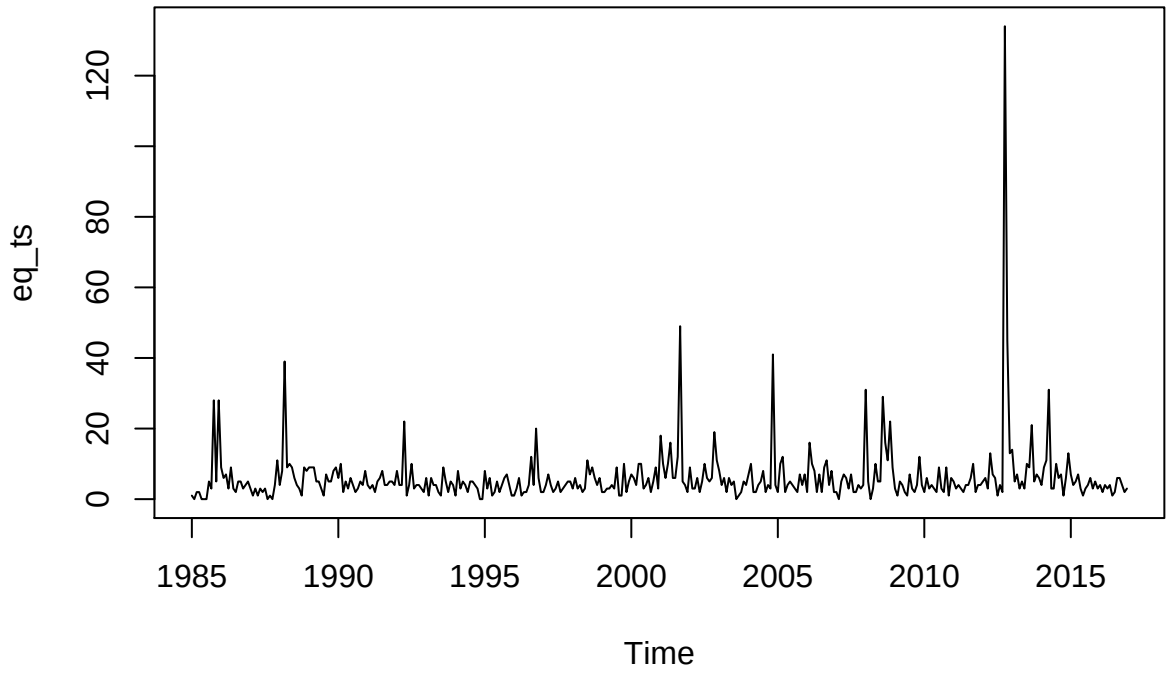


To assess the forecasting accuracy of the ARIMA model for earthquake events, we processed the original data into a monthly data from 1985 January to 2019 December. Subsequently, we computed the count of earthquakes with a magnitude greater than 4 for each month and year, resulting in a corresponding time series and we utilized the data from the first 32 years as the training set, reserving the data from the subsequent 3 years as the testing set.

Figure() illustrated this time series, and upon observation, it was evident that, with the exception of a few months characterized by a higher frequency of earthquakes exceeding magnitude 4, the curve generally fluctuated around a relatively constant level. The amplitude of these fluctuations was minimal, suggesting stationarity in the time series.

```
eq_data1 <- eq_data %>%
  select(date, magnitude) %>%
  filter(magnitude >= 4) %>%
  mutate(year_month = as.yearmon(date))

summary_data <- eq_data1 %>%
  group_by(year_month) %>%
  summarize(row_count = n())

summary_data$year_month <- as.Date(as.yearmon(summary_data$year_month))
```

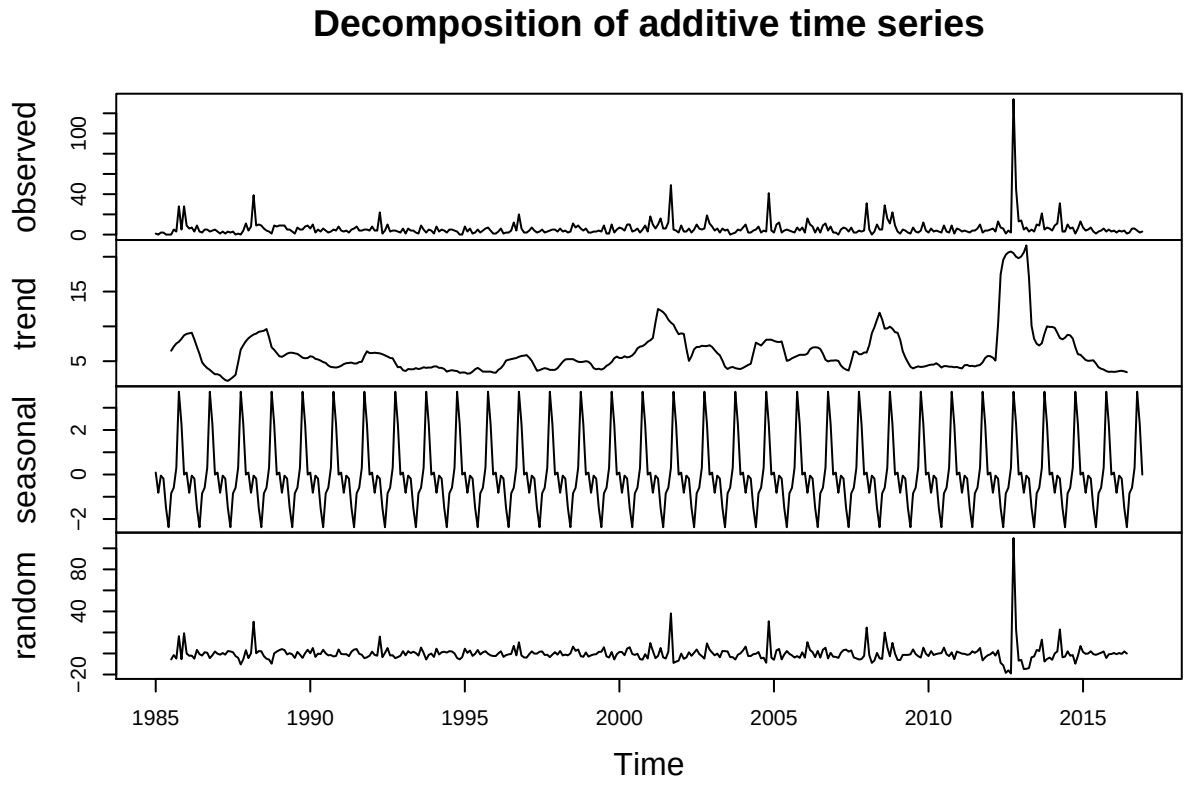
```r
all_months <- seq(as.Date("1985-01-01"), as.Date("2019-12-01"), by = "months")
all_months_data <- data.frame(year_month = as.Date(as.yearmon(all_months)))

summary_data <- all_months_data %>%
  left_join(summary_data, by = "year_month") %>%
  replace_na(list(row_count = 0))

eq_ts_all <- ts(summary_data$row_count, frequency = 12, start = c(1985,1))
eq_ts <- window(eq_ts_all, start=c(1985,1), end=c(2016,12))
eq_ts_test <- window(eq_ts_all, start=c(2017,1), end=c(2019,12))

plot(eq_ts)
```



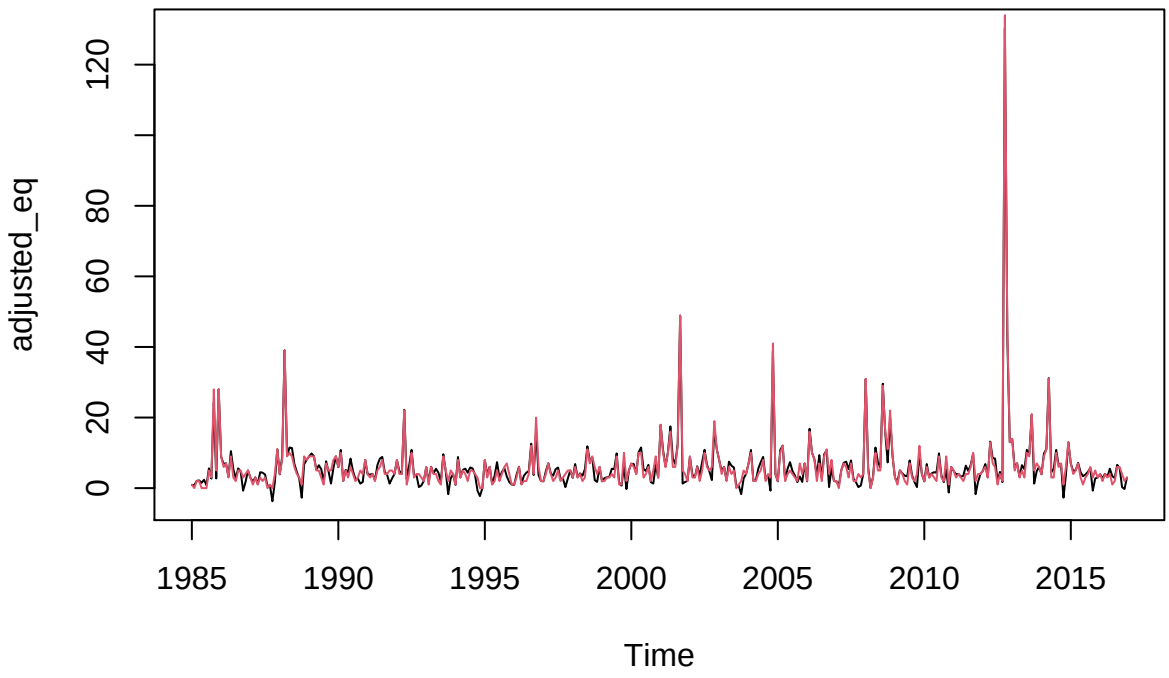
Next, we conducted a decomposition analysis on this time series. Figure() displayed, from top to bottom, the original time series plot, trend plot, seasonal plot, and residual plot. From the graphs, we observed that the fluctuation in the occurrence of earthquake with a magnitude exceeding 4 was not pronounced, maintaining a relatively stable state, except for a few sudden increases.

By subtracting the seasonal component from the original time series, we obtained a seasonally adjusted time series. The results are depicted in Figure(): the black curve represented the original time series, while the red curve represented the time series after removing the seasonal component. The small difference between the two curves indicated that the influence of seasonal variations on the frequency of seismic events with a magnitude exceeding 4 was minimal.

```r
decomposed_eq <- decompose(eq_ts, type="additive")
plot(decomposed_eq)
```




```
m_t <- decomposed_eq$trend #get the trend
s_t<- decomposed_eq$seasonal #get the seasonality
e_t<- decomposed_eq$random # get the random component

adjusted_eq <- eq_ts-decomposed_eq$seasonal
plot(adjusted_eq)
lines(eq_ts, col=2)
```



To validate our hypothesis about the lack of pronounced seasonal and trend changes, we applied a linear regression model to fit the data separately.

In the summary we saw that the p-value of the hypothesis test was 0.06847, bigger than 0.05, which meant we could not reject the null hypothesis $H_0 : \beta_1 = 0$. The same as the seasonal changes, because the p-value was equal to 0.06518, so we could not reject the null hypothesis $H_0 : \beta_1 = 0$. Therefore we could assume that all these three components were insignificant in the time series.

```
z_t<-(eq_ts-s_t) #remove the seasonality first to fit the trend
z_t<-data.frame(Index=index(z_t), Zt=c(z_t))
Regression<-lm(Zt~Index, data=z_t) #fitting a simple linear regression model w.r.t time
slope<-Regression$coefficients[["Index"]] #the parameter of the linear model
intercept<-Regression$coefficients[["(Intercept)"]] #the intercept of the linear model
summary(Regression)
```

```
##
## Call:
## lm(formula = Zt ~ Index, data = z_t)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.019  -3.504  -1.562   0.975 123.157
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -169.32506   96.00922  -1.764   0.0786 .
```

```
## Index          0.08767    0.04798   1.827   0.0685 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.685 on 382 degrees of freedom
## Multiple R-squared:  0.008663,   Adjusted R-squared:  0.006068
## F-statistic: 3.338 on 1 and 382 DF,  p-value: 0.06847
```

```
z_t1<-(eq_ts) #remove the seasonality first to fit the trend
z_t1<-data.frame(Index=index(z_t1), Zt1=c(z_t1))
Regression<-lm(Zt1~Index, data=z_t1) #fitting a simple linear regression model w.r.t time
slope<-Regression$coefficients[["Index"]] #the parameter of the linear model
intercept<-Regression$coefficients[["(Intercept)"]] #the intercept of the linear model
summary(Regression)
```

```
##
## Call:
## lm(formula = Zt1 ~ Index, data = z_t1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
##  -6.741  -3.588  -1.670   0.466 126.846
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -174.19000   97.48493  -1.787   0.0748 .
## Index          0.09010    0.04872   1.849   0.0652 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.819 on 382 degrees of freedom
## Multiple R-squared:  0.008874,   Adjusted R-squared:  0.006279
## F-statistic:  3.42 on 1 and 382 DF,  p-value: 0.06518
```

Before establishing an appropriate model, it was essential to conduct a stationarity test on the time series data. We posit the null hypothesis as a non-stationary time series, with the alternative hypothesis being 'stationary.' The results of the Augmented Dickey-Fuller (ADF) test yielded a p-value of 0.01, which was less than the significance level of 0.05. This provided sufficient evidence to reject the null hypothesis, indicating that the time series was stationary. Consequently, we could proceed to the next steps: determining the model order and fitting the model.

```
#ADF Test
adf_result <- adf.test(adjusted_eq)
```

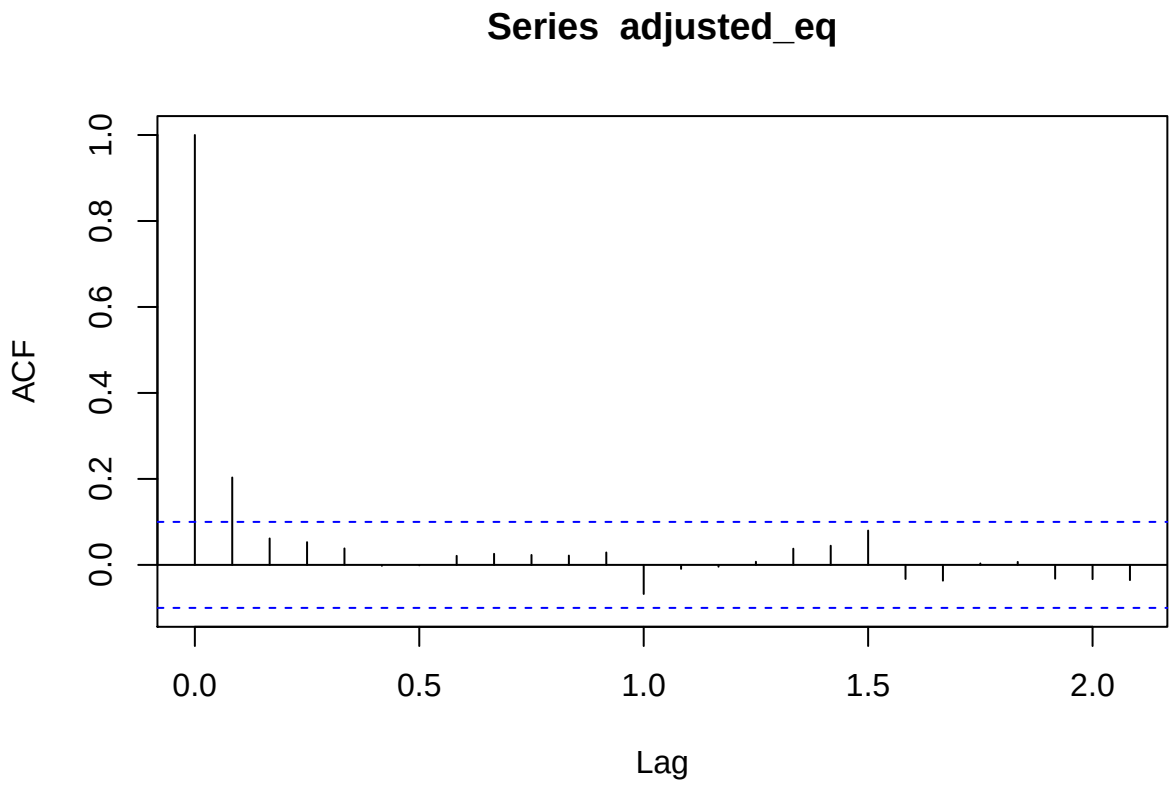
```
## Warning in adf.test(adjusted_eq): p-value smaller than printed p-value
```

```
adf_result
```

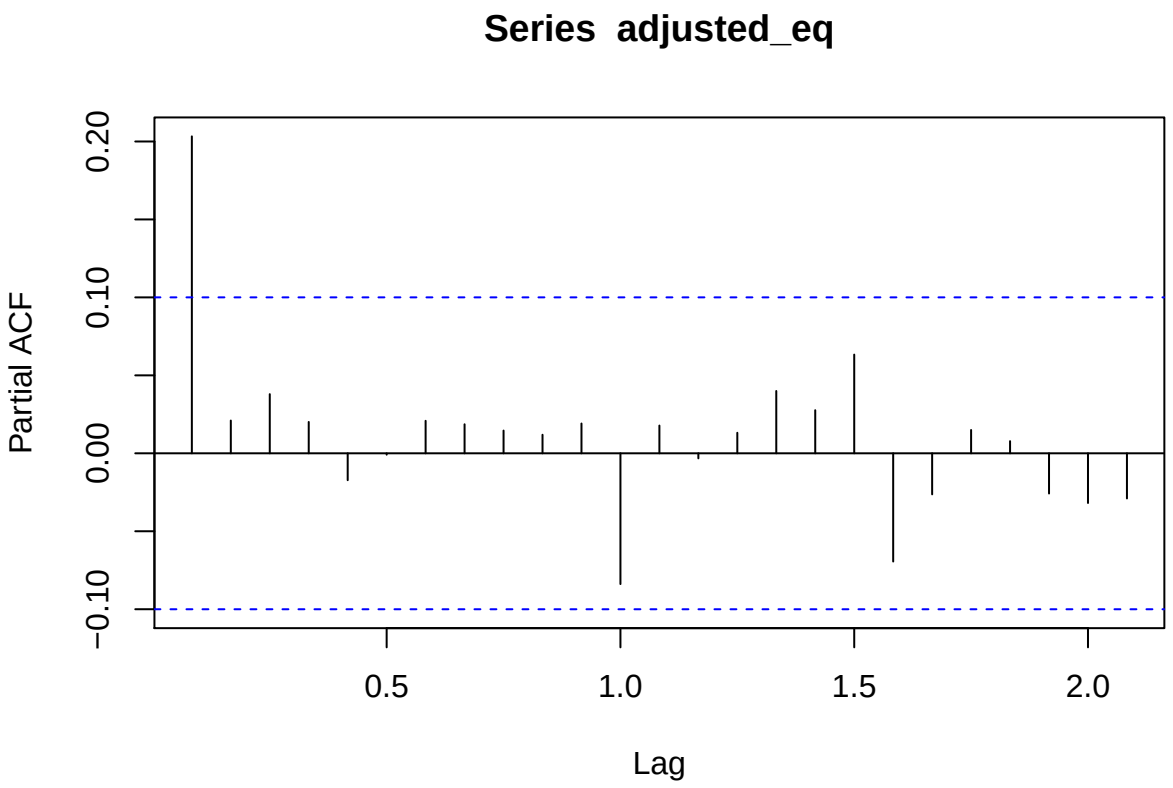
```
##
##  Augmented Dickey-Fuller Test
##
## data:  adjusted_eq
## Dickey-Fuller = -6.3105, Lag order = 7, p-value = 0.01
## alternative hypothesis: stationary
```

Generally, we determined the values of p and q in an ARMA model by observing the ACF and PACF plots. Figures() and () represented the ACF and PACF plots, respectively. It could be observed that the ACF was relatively high for the first two lags, surpassing the threshold line (blue line), and then gradually decrease. Thus, we inferred that the ACF was exhibiting a tailing pattern. On the other hand, the PACF cut off after lag 1 and subsequent PACF for all lags remained within the threshold line (blue line), indicating a truncated PACF. Consequently, we opted for an AR(1) model.

```
#ACF and PACF plot
acf(adjusted_eq)
```




```
pacf(adjusted_eq)
```



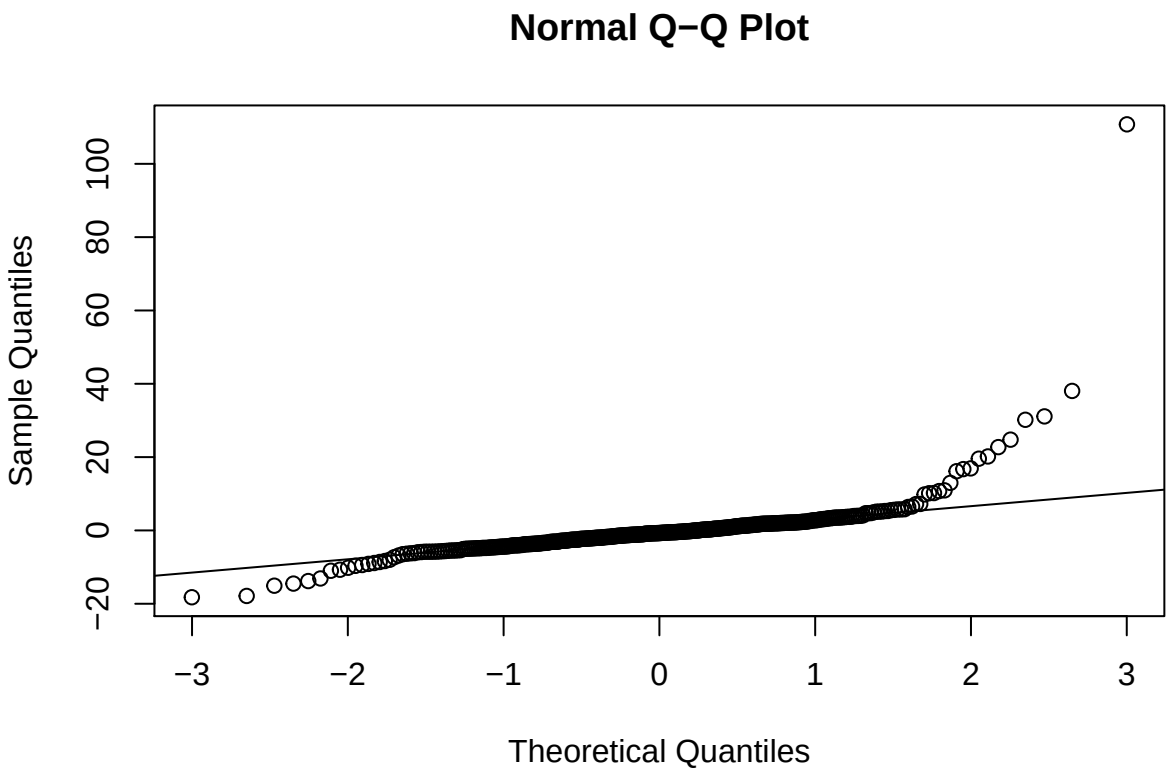

```
#fit an AR(1) model
model <- arima(e_t, order=c(1,0,0))
```

If a model was appropriate, the residuals should follow a normal distribution with a mean of 0, and for any lag order, the autocorrelation of residuals should be 0. In other words, the residuals of the model should exhibited independent and normally distributed behavior.

We generated a QQ plot for the residuals of the AR(1) model. As depicted in Figure(), the points on the QQ plot roughly aligned with a straight line, suggesting that the residuals approximately followed a normal distribution.

Another method to assess the adequacy of the model was the Ljung-Box test, which examined whether there was autocorrelation in the residual sequence. The result yielded a p-value of 0.8944, larger than 0.05. This indicated that the residuals had not passed the significance test for autocorrelation, suggesting that the autocorrelation coefficients of the residuals were close to zero. Therefore, the AR(1) model provided a good fit to the data.

```
#QQ plot of residuals
qqnorm(model$residuals)
qqline(model$residuals)
```



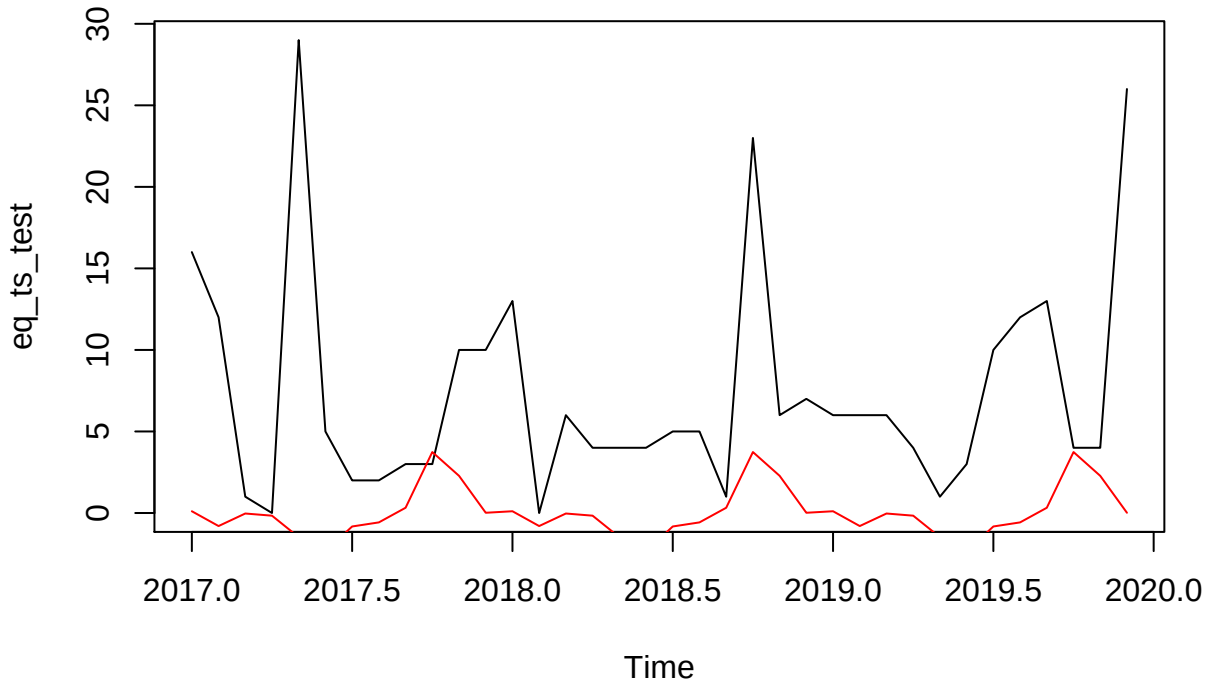

```
#Box-Ljung Test
Box.test(model$residuals, type="Ljung-Box")
```

```
##
##  Box-Ljung test
##
## data:  model$residuals
## X-squared = 0.017605, df = 1, p-value = 0.8944
```

Lastly we combined the predicted random component and the deterministic component together to predict the monthly number of earquakes which happened between 2017 to 2019 years and compared the result with the test data. In the figure(), the black line was the truth data and the red one was the prediction with the ARIMA model. It was obvious that there was a really significant difference between these two lines. The forecasting results from the ARIMA model exhibited a roughly similar trend to the actual data only for the period from May 2018 to December 2019. However, the magnitudes of the predicted values significantly deviated from the actual observations. Consequently, the forecasting performance of the ARIMA model was deemed unsatisfactory and did not provide meaningful insights.

```
#Prediction of error terms using the Box-Jenkins method
model_pred <- predict(model, n.ahead = 36)
Random_pred <- model_pred$pred  ## predicted future values

#predicted deterministic part:
Determ_pred<-s_t[1:36]
```

```r
#Combine the deterministic and random parts:
eq_ts_pred<-Determ_pred+Random_pred

plot(eq_ts_test)
lines(eq_ts_pred,col="red")
```



References

Amei, Amei, Wandong Fu, and Chih-Hsiang Ho. 2012. "Time Series Analysis for Predicting the Occurrences of Large Scale Earthquakes." *International Journal of Applied Science and Technology* 2 (7).

Nicolis, Orietta, Francisco Plaza, and Rodrigo Salas. 2021. "Prediction of Intensity and Location of Seismic Events Using Deep Learning." *Spatial Statistics* 42.

Ogata, Yosihiko. 1988. "Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes." *Journal of the American Statistical Association* 83 (401): 9–27.

ROSS, GORDON J., and ALEKSANDAR A. KOLEV. 2022. "SEMIPARAMETRIC BAYESIAN FORECASTING OF SPATIOTEMPORAL EARTHQUAKE OCCURRENCES." *The Annals of Applied Statistics* 16 (4): 2083–2100.