# STAT 403 Group 1 Final Project

Xiaoyan (Kelly) Peng, Danni Shi (Group leader)
Baichuan (Forrest) Zhang, Yi (Alan) Zhang

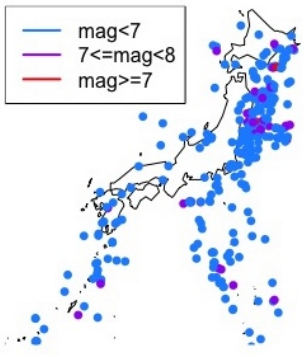
Spring 2017

## 1 Introduction:

Although recently many astronomers have been questioning the relationship between earthquakes and the movement of solar objects, people have never stopped testing and guessing, and the two most popular guesses are whether new / full moon causes earthquakes and whether solar / lunar eclipses cause earthquakes. The first guess is probably owing to the change of tidal force because of moon phase, and the latter guess has long been in myths and tales.

Our group chose data of earthquakes with 6.0 or higher Richter magnitude in Japan, a country located on the Pacific "Ring of Fire", from 1986 to 2016. Specifically, we studied the possible relationship between earthquakes and the positions of the Sun, the star of our solar system, and the Moon, the major satellite of planet earth. We also added data of all solar / lunar eclipses from 1986 to 2016. Here we only select data of main shocks and ignore data of after shocks or group shocks and we use this data to analyze those two guesses.

Our dataset, SolarSystemAndEarthquakes, was originally collected from the websites of USGS Hazards Program and calculated on NGINOV's website and was further arranged and shared on Kaggle.com.




the location and magnitude of earthquakes in Japan

# 2 Analysis: Japan

## 2.1 Our difficulties

At the first stage, we attempted to find the linear regression between earthquake magnitudes and all other variables. After excluding ineffective variables like earthquake latitude and longitude, we chose eight predictors: Moon phase value, Moon phase illumination, Sun longitude, Sun latitude, Sun azimuth, Moon longitude, Moon latitude and Moon azimuth. The summary shows that Moon azimuth whose p-value is 0.00337 is statistically significant. As we mentioned in introduction, we cleaned our raw dataset by deleting all aftershocks manually. Using the updated dataset, we find that p-values for Moon azimuth and Sun latitude are both less than 0.05.

Moreover, we used AIC and BIC to select the best model, which should be consistent with results of p-values. With updated dataset, AIC selects the model that "magnitude $\sim$ Sun longitude + Moon azimuth" while BIC shows that "magnitude $\sim$ Moon azimuth". The results of p-value and model selection are consistent. Again, using linear regression only among these three variables, p-values for Sun longitude and Moon azimuth are less than 0.05. So, we can say that the best model will be "magnitude $\sim$ Sun longitude + Moon azimuth".

```
Step:  AIC=-347.97
dat$earthquake.mag ~ dat$Sun.longitude + dat$Moon.azimuth

                            Df Sum of Sq    RSS     AIC
<none>                                    36.463 -347.97
+ dat$Moon.latitude          1   0.26654 36.197 -347.48
+ dat$Moon.longitude         1   0.13433 36.329 -346.73
+ dat$Sun.latitude           1   0.12437 36.339 -346.67
+ dat$MoonPhase.value        1   0.11840 36.345 -346.64
+ dat$MoonPhase.illumination 1   0.01711 36.446 -346.07
+ dat$Sun.azimuth            1   0.00488 36.459 -346.00
- dat$Sun.longitude          1   0.74119 37.205 -345.85
- dat$Moon.azimuth           1   1.61729 38.081 -341.08
>
```

Figure 1: AIC model

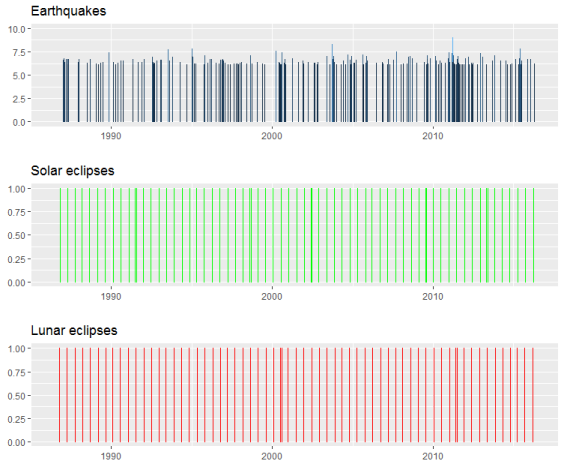
```
Step:  AIC=-339.2
dat$earthquake.mag ~ dat$Moon.azimuth

                            Df Sum of Sq    RSS     AIC
<none>                                    37.205 -339.20
+ dat$Sun.longitude          1   0.74119 36.463 -338.00
- dat$Moon.azimuth           1   1.58839 38.793 -335.96
+ dat$Moon.latitude          1   0.26559 36.939 -335.35
+ dat$MoonPhase.value        1   0.13727 37.067 -334.64
+ dat$Moon.longitude         1   0.12783 37.077 -334.58
+ dat$Sun.latitude           1   0.08995 37.115 -334.38
+ dat$MoonPhase.illumination 1   0.02123 37.183 -334.00
+ dat$Sun.azimuth            1   0.00000 37.205 -333.88
>
```

Figure 2: BIC model

However, the estimated coefficients for Sun longitude and Moon azimuth are 0.0005850 and 0.0009370, which are really close to 0. Although p-values are small, the estimated coefficient still suggests that there is no linear relationship or extremely weak linear relationship among three variables.

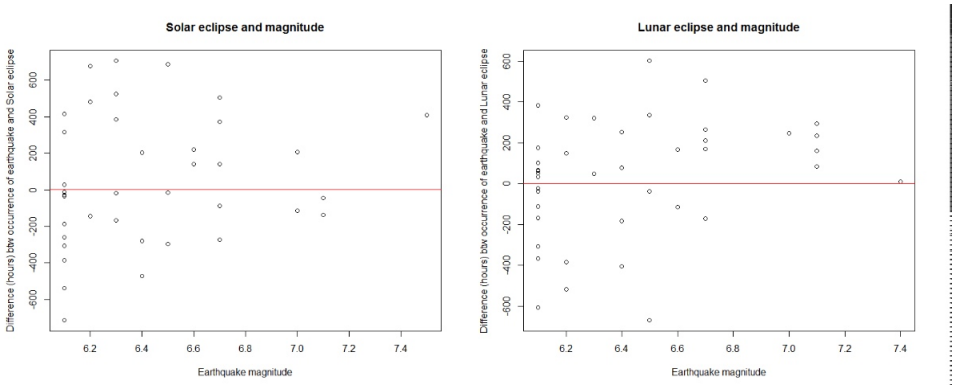
## 2.2 Solar and Lunar Eclipses

xAs recorded in Matthew's Gospel, when Jesus Christ was crucified, a solar eclipse, which might later stir an earthquake, occurred. Here we try to analyze their possible relationship so we try to find the correlation between earthquake magnitude and the difference between solar / lunar eclipse time and time of earthquakes.



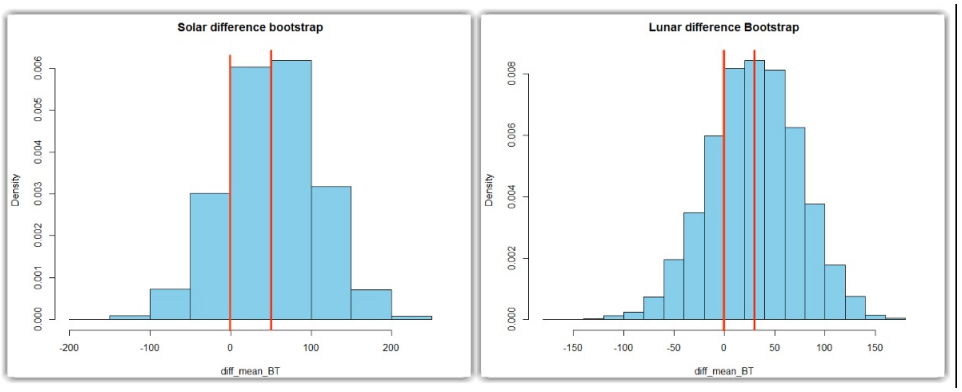

We only select data of earthquakes within 600 hours (25 days) of an eclipse, and if there was relationship between eclipses and earthquakes, we would expect points surrounding the red horizontal line with y-intercept is 0. So, we made following hypothesis testing:

$$H_0 : \textbf{mean}(\text{diff}) = 0$$
$$H_1 : \textbf{mean}(\text{diff}) \neq 0$$




scatterplot of magnitude and time difference between an earthquake and an eclipse

To further measure the accuracy of sample estimates, we used empirical method to generate random sampling with replacements. We bootstrapped the means of differences for both Solar and Lunar cases 10000 times and get the following histogram. Then we applied t-test to test the null hypothesis. For both lunar and solar cases, the p-values of t-tests are so small that we can reject our null hypothesis.

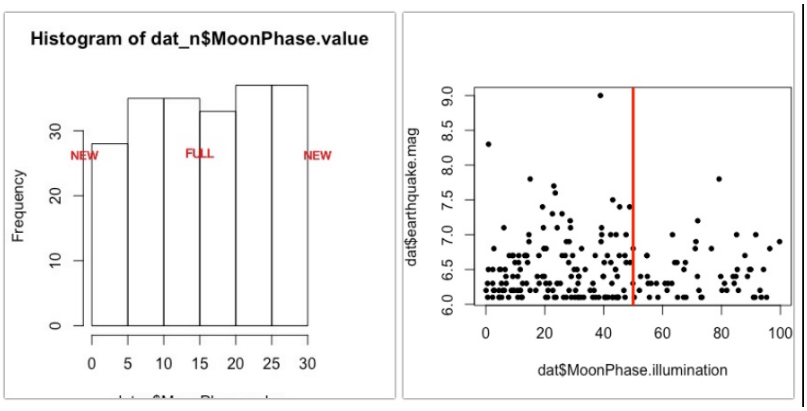


## 2.3 Full / New Moon Part 1: General Analysis

To investigate the relationship between moon phases, especially during full/new moon, and occurrence of earthquake, we need to test two hypothesis. First we would analyze whether earthquakes occur more often during New/Full moon. By plotting the histogram of moon phase value, we made the hypothesis that

$$H_0\text{: Occurrence of earthquake} \sim \mathsf{Uni}(n, 0, 29.53)$$
$$H_1\text{: Occurrence of earthquake} \nsim \mathsf{Uni}(n, 0, 29.53)$$

Then we used the KS test to verify our null hypothesis. The large p-value (0.9997616) cannot reject null hypothesis and suggests that occurrence of earthquake could be randomly distributed.




left: histogram of moon phase value; right: scatterplot of illumination and magnitude

Second, we want to test if the magnitudes of earthquakes increase during full / new moon, and here we analyze the relationship between earthquake magnitude and moon phase illumination (i.e., how bright the moon is (by percentage). When it's full moon, it goes to 100 and when it's new moon, it goes to 0).

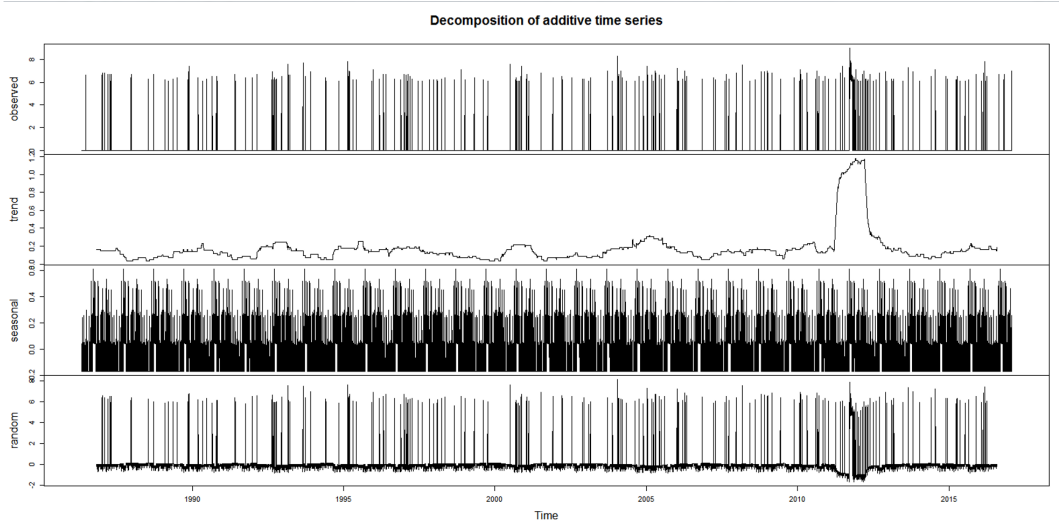
Ideally, for illumination within 0 to 50, we would expect a negative (linear) regression and a positive one within 50 to 100. Here we used residual bootstrap and for the bootstrap coefficients, only 0.14% and 11.4% bootstrap samples satisfy and have $< 0.05$ p-values for illumination percentage $< 50$ and $>= 50$ respectively. More over, we find it very hard to tell if there is any clear positive or negative regression for both parts.

## 2.4 A Time Series Model

Time series analysis accounts for the internal structures (such as auto-correlation, trend or seasonal variation) that observations taken over time may have. It not only offers insights into the underlying forces and structure that produced the observed data, but also eases future model fitting and forecasting process. This trait is highly valuable to our study, because our research interest is in determining the causes of earthquake occurrence, not merely in modelling and/or forecasting.

In our project, we use time series methods mainly as an analytical tool for validating the pre-proposed potential causes of earthquake occurrence. Since time series analysis

requires an observation value at each time step, we manually fill in the corresponding value of the dates with no earthquake of above 6.0 Richter as 0. After reformatting the data, we first apply the "TTR" package in R (which uses block bootstrap method) to decompose the historical earthquake data into three components, namely the overall growth trend, the seasonal period, and the random noise, and then compare the period of fluctuation of each the pre-proposed explanatory variables (including their mutual combinations after mathematical transformation) with our "seasonal period" acquired by trend decomposition.




One of the best-performing models, comprised of several explanatory variables, in our analysis is discussed in the following sub-section.

To fully explore the power of time series methods, we also attempt to use the time series model to predict future earthquakes, but that gives us a somewhat unreasonable result. The main reason is related to our aforementioned data cleaning approach; the details are discussed in Section 4.2.3.

## 2.5 Full / New Moon Part 2: Tidal Force

Given previous study, solar movements has limited effect on earthquake, but it is widely known the movements of sun and moon have effect on tidal force on earth. So we try to calculate the tide force with variables in our dataset, specifically, sun azimuth and moon azimuth. Tidal force contains vertical force $f_v$ added by sun and moon, and horizontal force $f_h$. The formulas are:

$$f_v = f_{mv} + f_{sv}; \qquad f_h = f_{mh} + f_{sh}$$

Other force of decomposition have names which follow specific pattern. For example, $f_{mh}$ denotes the Moon component of horizontal tide force. For each force of decomposition, use the formula below: ($G$: gravitational constant, $M_s$: mass of Sun; $D_s$: distance between Sun and earth; $\alpha$: angle between north pole, center of the earth, and the Sun; $\beta$: angle between earth and Moon)

$$f_{sh} = \frac{GM_s}{D_s^2}\left(\frac{sin\alpha}{(1-2\frac{R}{D_s}cos\alpha+\frac{R^2}{D_s^2})^{\frac{3}{2}}} - sin\alpha\right)$$

$$f_{sv} = \frac{GM_s}{D_s^2}\left(\frac{cos\alpha - \frac{R}{D_s}}{(1 - 2\frac{R}{D_s}cos\alpha + \frac{R^2}{D_s^2})^{\frac{3}{2}}} - cos\alpha\right)$$

$$f_{mh} = \frac{GM_m}{D_m^2}\left(\frac{sin(\beta)}{(1 - 2\frac{R}{D_m}cos(\beta) + \frac{R^2}{D_m^2})^{\frac{3}{2}}} - sin(\beta)\right)$$

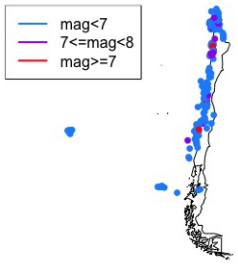
$$f_{mv} = \frac{GM_m}{D_m^2}\left(\frac{cos(\beta) - \frac{R}{D_m}}{(1 - 2\frac{R}{D_m}cos(\beta) + \frac{R^2}{D_m^2})^{\frac{3}{2}}} - cos(\beta)\right)$$

A linear regression analysis between earthquake and both new forces shows quite different result, which makes the decomposition of vertical and horizontal direction important. The result is that regression for earthquake magnitude and vertical force failed(p-value = 0.48). However, modeling with magnitude and horizontal tide force is successfully constructed(p-value = 0.055). Even though the slope is relatively small(around 0.08), it is acceptable, because when $f_{sh}$, $f_{sv}$, $f_{mh}$, $f_{mv}$ is calculated, the term $\frac{GM}{D^2}$ was omitted for simplicity, because it is assumed to be a constant. Therefore, the slope can still be influential for horizontal force. So, the result is significant.

To sum up, in theory, the relative position of Moon and Sun can influence tide force, and tide force can further trigger earthquake. The mantle layer(the one below earth crust) is soft enough to assume that position of Sun and Moon have influence in deep underground earth, which may cause plate movement. However, it would be hardly accepted by scholars that we can "predict" earthquake based on relative position of earth, Sun and Moon. Otherwise, thousands of lives will be saved each year around the world. Nevertheless, there definitely exists correlation in between.

# 3 Comparison: Meanwhile in Chile

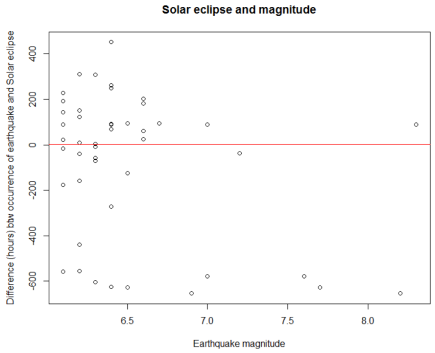
Chile is another country located on the Pacific "Ring of Fire". Like Japan, Chile also suffers from hundreds and thousands of earthquakes every year, but they are mainly caused by volcano activities. Here we also solely focus on data of main shocks as well.

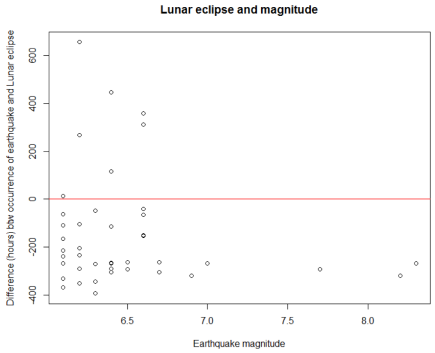


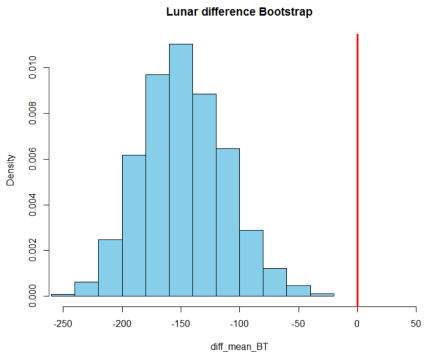

locations of earthquakes and magnitudes

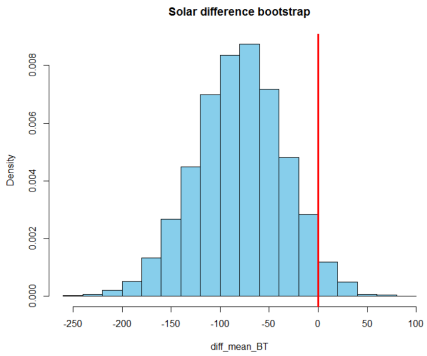
## 3.1 Solar and Lunar Eclipses

For Chile's case, we used the empirical bootstrap and t-test to test the same hypothesis we made in Japan's case. From the following bootstrap histogram, we see that there's even nothing gather around 0 in Lunar Eclipse case and only a few in Solar Eclipse case. Due to the very small p-value for t-test, we reject the null hypothesis and can conclude that there's no valid relationship between Solar/Lunar Eclipse and earthquake.



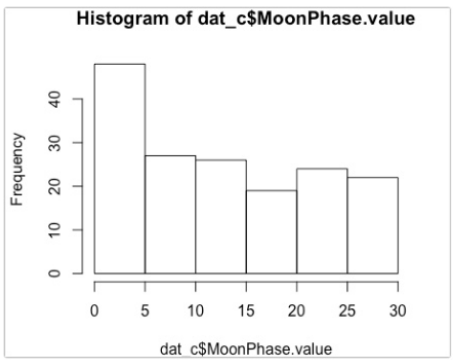



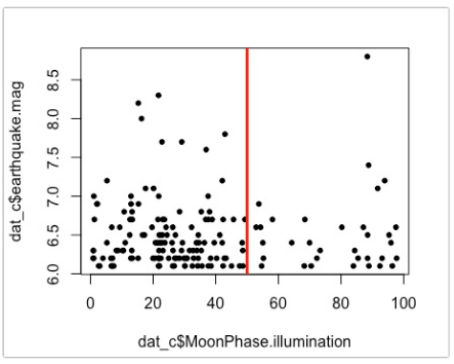




## 3.2 Full / New Moon: General Analysis

Similarly here we first test whether earthquakes occur more often during new / full moon. We do a KS test (with null hypothesis $H_0$ that the distribution for moon phase follows a uniform distribution `runif=(n,min=0, max=29.53)`). Since the p-value here is large (0.6121276), we fail to reject the null hypothesis as well.



Second we analyze the relationship between earthquake magnitude and moon phase illumination with the same expectations. Here only 1.5% and 15.06% bootstrap coefficients have $< 0.05$ p-values for illumination percentage $< 50$ and $>= 50$ respectively.

# 4 Conclusion & Discussion:

## 4.1 Conclusion:

At first it was hard to find possible relationship (linear) regression among earthquake magnitude and all other quantitative variables of sun and moon. After interpreting AIC, BIC and popular guesses, we decided to focus on the relationship between earthquakes and moon phase, eclipses and tidal force (caused by solar movements).

Based on our analysis of data of main shocks in Japan, we use KS test to find that the occurrence of earthquakes may not depend on moon phase and nor does the magnitude. Also, by using goodness of fit test and empirical bootstrap to find the relationship between earthquake occurrences and magnitudes and the time differences between an earthquake and an eclipse, we find that there may not be a clear relationship. By using the same methods to analyze data of main shocks in Chile, we find at least these results are consistent.

Then by applying sun and moon azimuths in our data, we try to find the tidal force caused by the solar movements, but we fail to find any significant fitting model. We also use block bootstrap to find a time series model, but it's not guaranteed to be a fitted one.

## 4.2 Drawbacks:

### 4.2.1 Data selection:

As we mentioned, we select data of "main shocks" manually, by which we mean that we only select data of earthquakes happened first, and exclude data of earthquakes within 2 days and $\pm 1.5$ degrees in latitude / longitude on the land and $\pm 3.5$ in the sea. However, it's usually hard to categorize fore shocks, main shocks and after shocks. For instance, an aftershock may occur months after the main shock and not always at the same location; before a main shock, there could be multiple fore shocks; recently some astronomers claim that an after shock could be stirred by a change in tidal forces. So it's possible that we excluded some true signals and remained some noises in our updated dataset.

On the other hand, we only investigate data of two countries on the "Ring of Fire", and our results may not be representative for earthquakes solely caused by collision of plates, such as earthquakes occurred near Tibet.

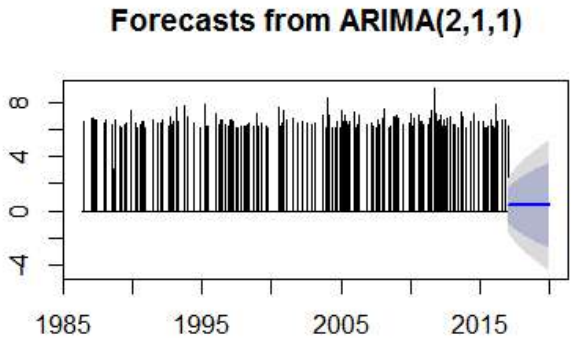
### 4.2.2 Tidal Force: About the Contingency Table

Another place to improve is that, when doing tide force analysis, we find that a linear regression might not fully interpret the complicated physical model. Hence doing a contingency table may help study the dependency. The two attributes involved in contingency table can be magnitude and tide force(vertical or horizontal). It is reasonable to divide the earthquake magnitude as "huge" if the level of magnitude is greater than median and "small" if less. Similarly, divide the force as "huge" and "small". By doing two chi-square tests in R (magnitude with both forces), neither of them indicates statistical significance. However, if the data includes all earthquake (with aftershocks), those two chi square tests revealed non-negligible significant(with p-value of 0.078 and 0.099). Even though it is not significant enough, it is small enough to indicate rareness and curiosity. A reasonable explanation for using the whole data set(include aftershock) is that after each major earthquake, the earth plate becomes active, which makes the effect caused by sun and moon more influential.

### 4.2.3 Time Series Forecasting: A Pitfall in Data Cleaning

Even though time series methods effectively help us identify and validate the most appropriate explanatory models among the many potential candidates we found in a wide range of geo-scientific papers, it does not function satisfactorily when forecasting future earthquakes based on the historical data we use throughout the study.

The graphic below illustrates the forecast of future earthquakes based on the time series model, with 80% and 95% Confidence Interval.




This does not make much sense in the context of our study. The range from 0.0 to 6.0 is not used by our dataset, but it is included in the range of forecast. Therefore the predicted magnitude within the range between 0.0 to 6.0 has no actual meaning. To emphasize, this does not mean that we would expect earthquakes of magtitude between 0.0 and 6.0 in the near future; the range between 0.0 and 6.0 is just completely undefined in the context of our study.

To solve this issue, ideally we would need to base our analysis on a comprehensive dataset that includes the record of all earthquakes of all magnitude, so that the defined

range of prediction would include all positive real numbers. But this is practically impossible, since it is very difficult to detect earthquakes of very small magnitude.

## 4.3 After words:

This topic did not start to be fully discussed until 1900s, and is still an intriguing one in the field of earthquake prediction, and for the four of us, this project is our own attempt. Due to many reasons, (lack of geophysical knowledge, lack of data, etc.) we fail to fit a significant model, but we highly treasure the process of working together and analyzing the data. We believe as human make further exploration about the cause of earthquakes, the prediction can be more and more accurate in the future.

# 5 Reference:


1. USGS (United States Geological Survey) Hazards Program: https://earthquake.usgs.gov/

2. Kaggle, Earthquakes and Solar System Objects, https://www.kaggle.com/aradzhabov/earthquakes-solar-system-objects, April, 2017.

3. Klotz, Otto, *Earthquakes, Phases of the Moon, Sub-lunar and Sub-solar Points*, Jour Royal Astronomy Society, Canada, 1914, pp. 273-281

4. Chiou, L., *The Association of the Moon and the Sun with Large Earthquakes*, https://arxiv.org/pdf/1210.2695.pdf, Oct 9, 2012

5. Hsu, Kuo-cheng (20030800) *Tide force for Sun and Moon,Combined category of Digital Collection and Learning*, http://catalog.digitalarchives.tw/item/00/4a/5a/6a.html accessed on 2017/06/01