Capstone Project Report

Are we doing better at reducing the total loss in large earthquakes?

Chen Chen

01/17/2018

Introduction

Earthquakes, one of the most damaging natural disasters, take away hundreds to thousands of lives and houses worldwide each year. While it is difficult to predict when the next earthquake will hit, the earthquake prone areas can take actions to be more prepared if such disasters happen. Through this project, I want to analyze earthquakes for the past 50 years to investigate (1) the factors that contributed to high fatalities/damages in large earthquakes and (2) the correlation between large earthquake fatality/damages and a country's economic situation. The questions that I attempt to answer through this project are:

1) overall, are we doing better at reducing damages and fatality over the past 50 years?
2) are the total deaths/damages correlated with the magnitude/depth/location of the earthquakes?
3) is there a correlation between a country's GDP/population and the total deaths/damages?

Explanation of the Dataset

The dataset used in this study is downloaded from NOAA significant earthquake database, which contains damaging earthquakes from 2150 B.C. to the present. Earthquakes between 1967 and 2017 were downloaded for this analysis. This dataset contains important information of the earthquake date, location, depth, magnitude, total deaths, and total damages. A country's economic status for the past 57 years, represented by GDP and population were downloaded from the World Bank database.

The difficulty of analyzing the dataset is how to handle the missing values. For instance, 1222 out of 2012 earthquakes do not have the total deaths count, and 1691 out of 2012 earthquakes do not list the total damage values. It is unclear if the empty field represents unavailable data or there were no damage or deaths in these earthquakes. However, considering that the downloaded earthquake data were from the past 50 years when global earthquake recording networks were in place and documentations of these earthquakes were relatively complete, it is very likely that these missing fields represent no deaths or damages.

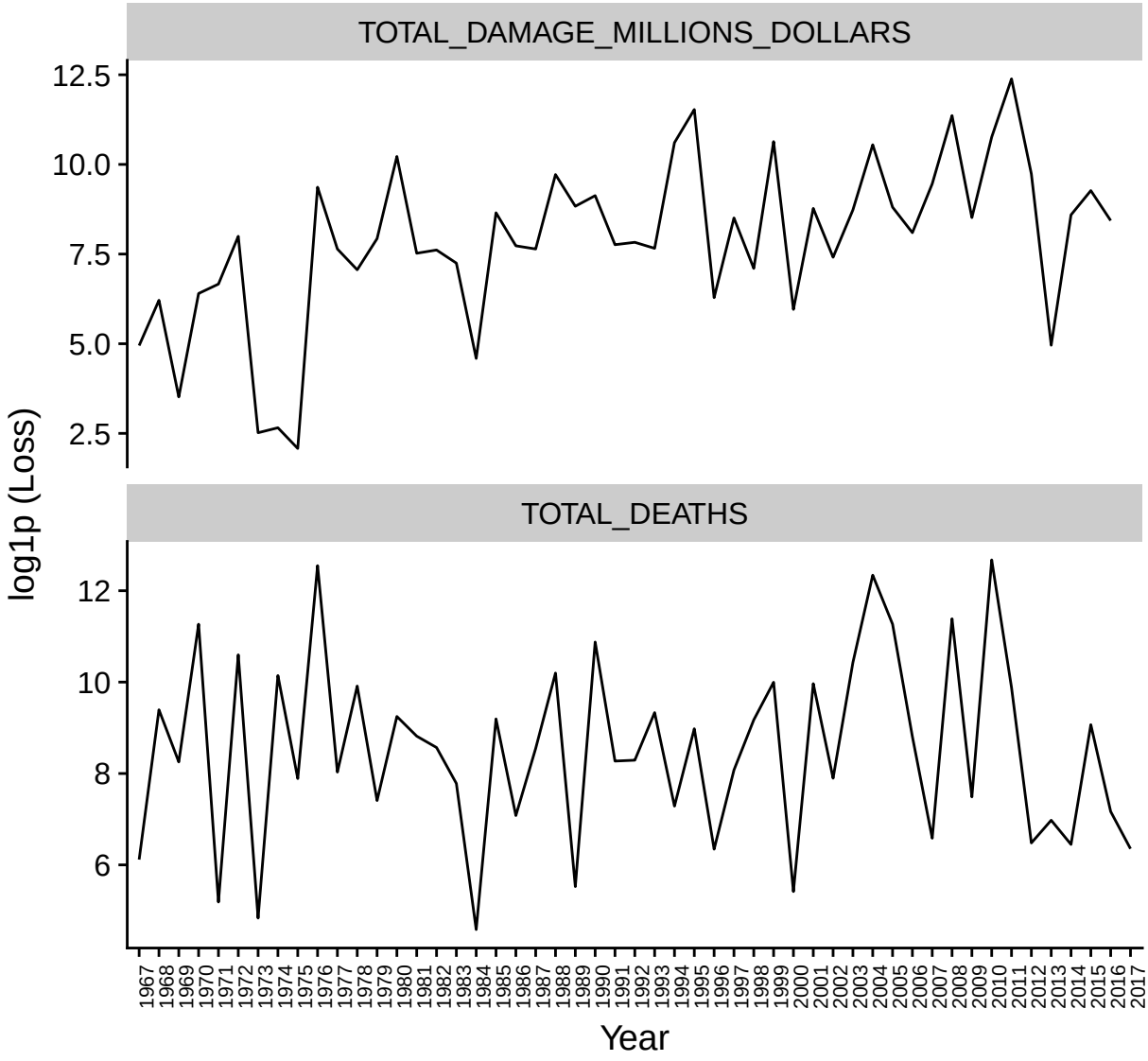
One limitation of this dataset is the absence the local population density information at each earthquake location. While a country's population is an indication of how dense the population is, the distribution of population varies from area to area. In addition, large countries will have great values of population, but the earthquakes may occur along faults that are far away from the the densely populated areas. Another limitation is the lack of information on building code/style at each earthquake location. One would imagine that areas with stronger buildings will have fewer deaths/damages. After all, earthquakes don't kill people, but buildings do. Because of the lack of information on the local population density and building code, we cannot quantify the correlation between these factors and the total deaths/damages.

To clean up the data, I converted the date and time of the earthquakes from characters to date in R. Using the "dplyr" and "tidyr" packages, I re-arranged the earthquake data by countries and earthquake variables that were used in this analysis. I also collapsed the GDP and population data into Year-Value pairs, in order to john the population/GDP data with with the earthquake data. There are ".." values in the GDP and population data, which were replaced with NA values.

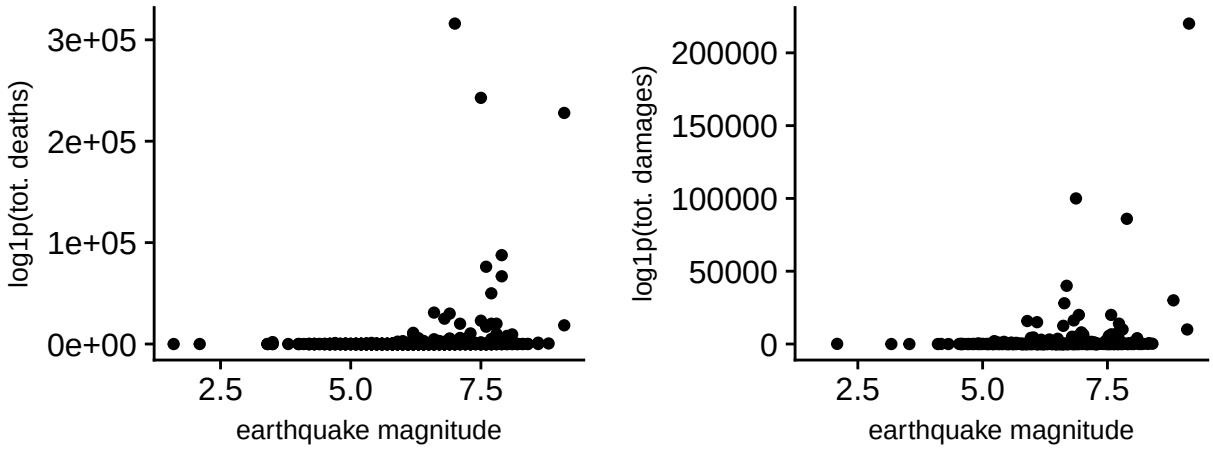
Results

Visualization of the Data

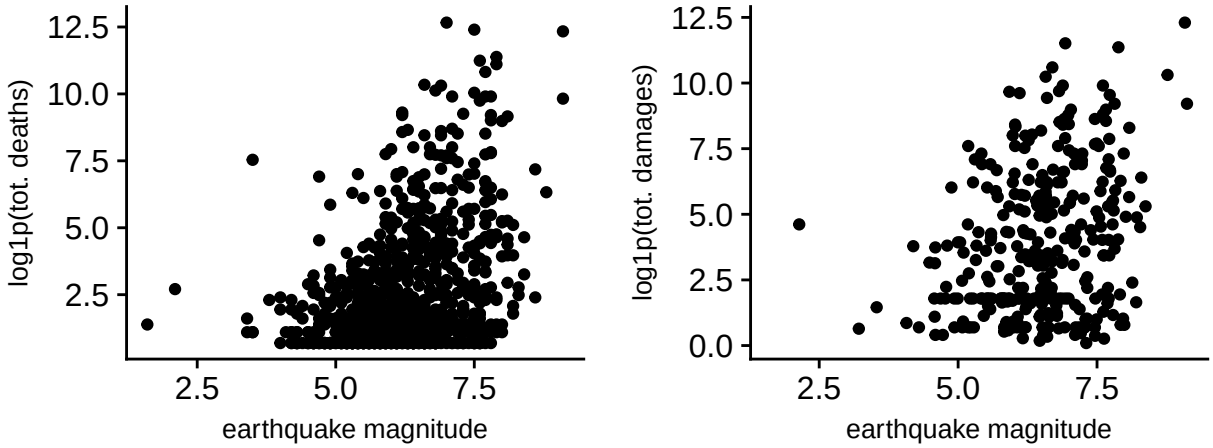
The "ggplot2" package was heavily used to generate the plots in this project. Plotting the total deaths/damages resulted from these significant earthquakes over the past 50 years reveal no decrease in the total deaths/damages in large earthquakes, which suggests that we are not doing better at reducing the total loss in large earthquakes.



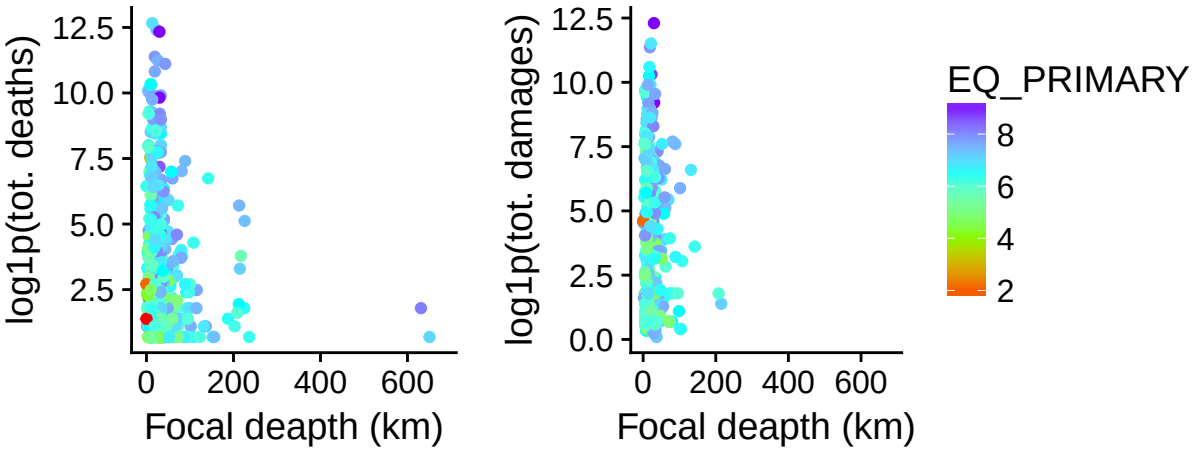
First, I plotted the total deaths and total damages VS earthquake magnitude. A few data points with very high fatality and damages skewed the plots and made the rest of the data points very clustered. So I applied log1p() to the total deaths and total damages before plotting, because log scale allows a large range to be displayed without small values being compressed down into bottom of the graph.



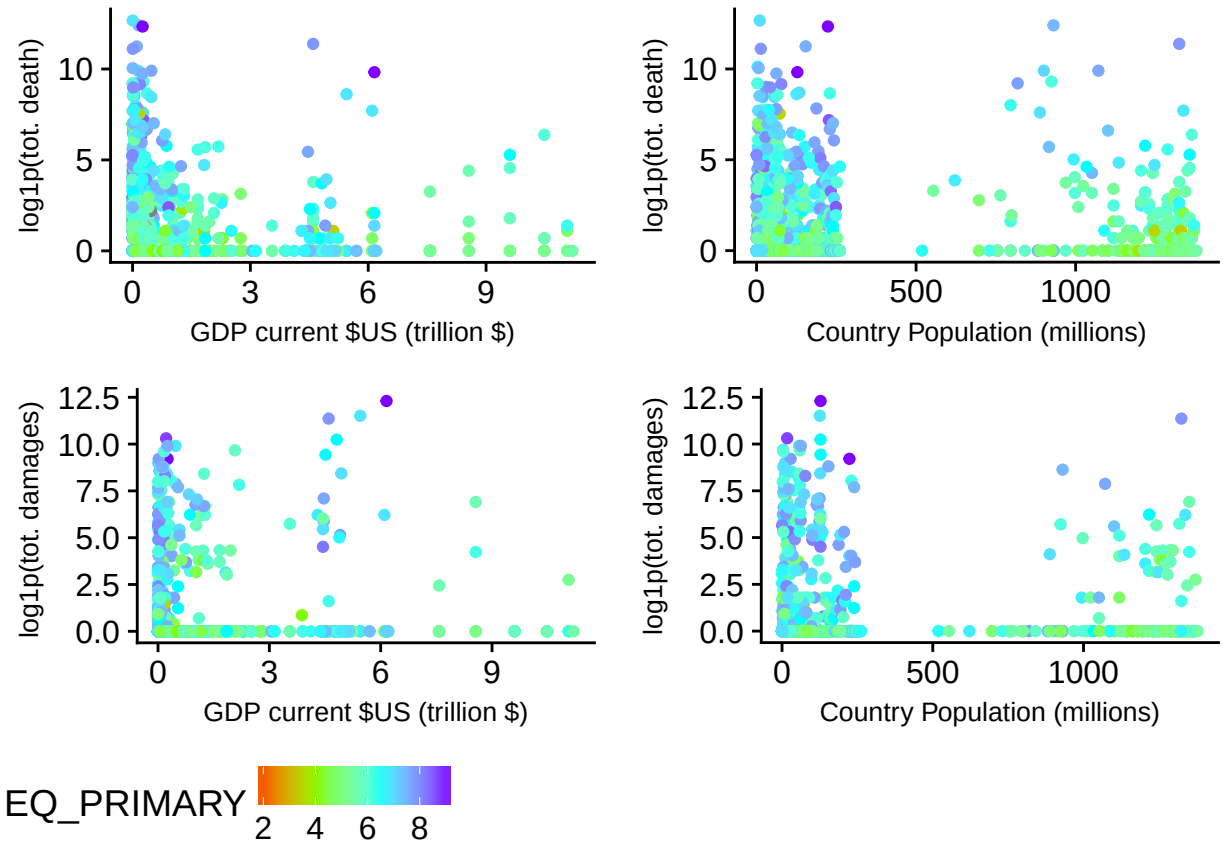
There is a positive correlation of earthquake magnitude and the total death/damages in an earthquake. As expected, more people died and more damages occurred in large earthquakes. Instead of a linear correlation, the distribution of the data points is more spreaded out.



There is a weak correlation between focal depth and total deaths. More deaths are correlated with shallower earthquakes. There is usually no death in very deep earthquakes (> 300 km), except for one earthquake in Peru that occurred in 1970 with 1 death on record, and the 1994 M8.2 earthquake in Bolivia that killed 5 people. No clear linear correlation is observed between focal depth and total damages. But it is interesting to see that most damages are related to earthquakes that are shallower than 100 km depths.



Another question I am interested in is if a country's population or its GDP value is correlated with the total loss in an earthquake. I plotted the total deaths/damages as a function of countries' GDP and population. No clear correlation between the total loss and a country's population can be identified in this dataset. The GDP data, however, seems to show that countries with high GDP values were subject to fewer deaths and less damages in large earthquakes.



Machine Learning Methods Applied to the Data

For this exercise, I applied two types of machine learning algorithms and both are supervised algorithm. I first ran linear regression analysis on my dataset, because I wanted to explore if the variable that I want to predict (total deaths/damages) varies as a function of the earthquake variables (such as depth, magnitude, location(lat,lon), and intensity). But the high residual and low R-squared value in the summary of this model suggest a bad fit.

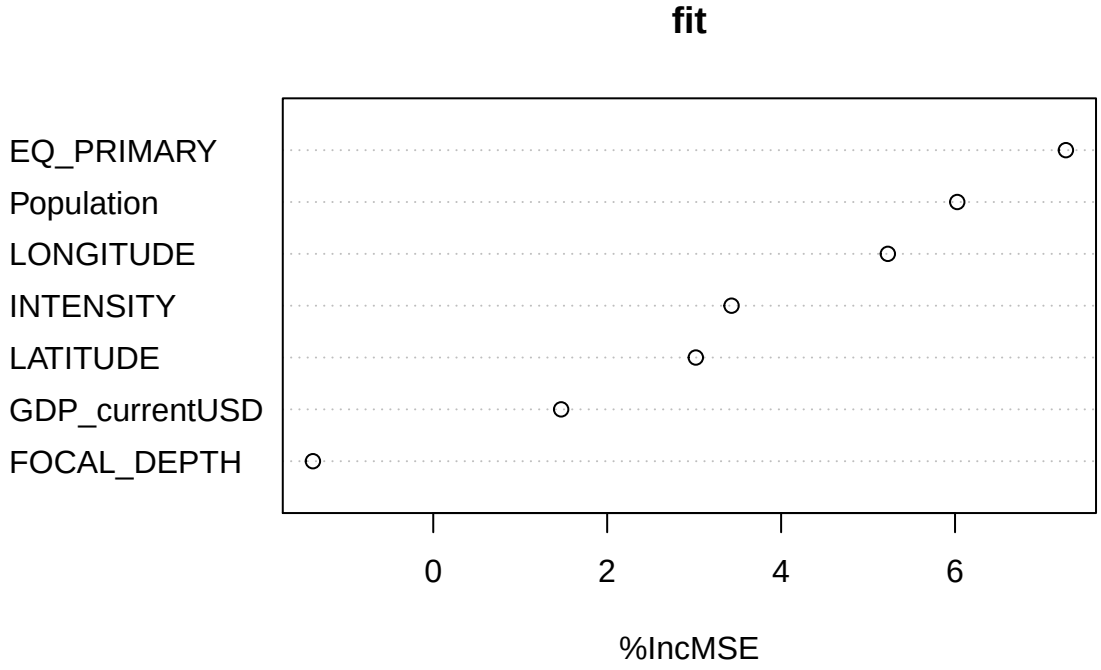
```
##
## Call:
## lm(formula = TOTAL_DEATHS ~ FOCAL_DEPTH + EQ_PRIMARY + INTENSITY +
##     LATITUDE + LONGITUDE, data = df_eq_gdp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -7037  -2475  -1083    490 235468
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -13720.361   3552.034  -3.863 0.000125 ***
## FOCAL_DEPTH     -6.499     14.024  -0.463 0.643229
## EQ_PRIMARY    1157.980    541.339   2.139 0.032837 *
## INTENSITY     1026.761    331.854   3.094 0.002068 **
## LATITUDE        18.136     20.368   0.890 0.373607
## LONGITUDE        4.021      4.932   0.815 0.415196
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11600 on 591 degrees of freedom
##   (1415 observations deleted due to missingness)
## Multiple R-squared:  0.03851,    Adjusted R-squared:  0.03038
## F-statistic: 4.734 on 5 and 591 DF,  p-value: 0.0002999
```

I also included the economic status of a country represented by the GDP and population in the regression analysis. Inspecting the summary of this model suggests a slight increase in the performance, but it still was not a good fit model.

```
##
## Call:
## lm(formula = TOTAL_DEATHS ~ FOCAL_DEPTH + EQ_PRIMARY + INTENSITY +
##     LATITUDE + LONGITUDE + GDP_currentUSD + Population, data = df_eq_gdp)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -14328  -2789   -673   1040 228473
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -1.815e+04  5.240e+03  -3.464 0.000593 ***
## FOCAL_DEPTH     5.888e-01  1.944e+01   0.030 0.975850
## EQ_PRIMARY      1.646e+03  8.395e+02   1.960 0.050704 .
## INTENSITY       1.098e+03  5.019e+02   2.187 0.029381 *
## LATITUDE        1.565e+01  3.117e+01   0.502 0.615937
## LONGITUDE      -3.968e+00  8.344e+00  -0.476 0.634703
```

```
## GDP_currentUSD  3.004e-10  9.400e-10   0.320 0.749451
## Population      8.406e-06  2.545e-06   3.303 0.001049 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13820 on 376 degrees of freedom
##   (1628 observations deleted due to missingness)
## Multiple R-squared:  0.07491,    Adjusted R-squared:  0.05769
## F-statistic:  4.35 on 7 and 376 DF,  p-value: 0.0001188
```

These tests show that simple linear regression analysis did not work well for this dataset. So I appealed for a more advanced algorithm, the Random Forests. Random Forest performs regression analysis by constructing multiple decision trees at training time and outputting the mean result of the individual trees. Using the "randomForest" package in R, I ran this algorithm on my data. The large mean of squared residuals and the negative value for the % of variance explained by this model indicates that random forest method was not adequate to define a good fit model that relates the total deaths/damages to the earthquake variables.

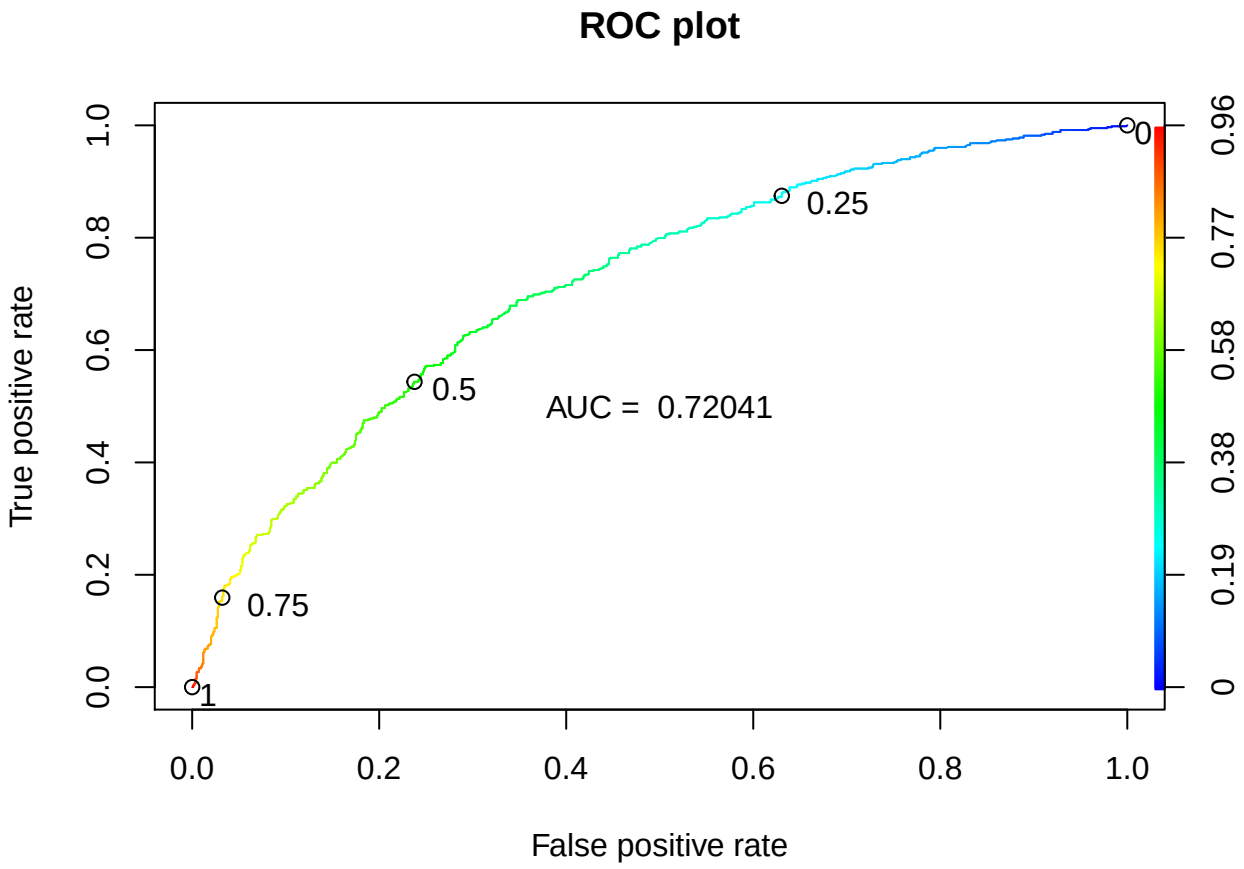



```
##
## Call:
##  randomForest(formula = TOTAL_DEATHS ~ FOCAL_DEPTH + EQ_PRIMARY +      INTENSITY + LATITUDE + LONGITU
##                Type of random forest: regression
##                      Number of trees: 2000
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 210667716
##                     % Var explained: -4.21
```

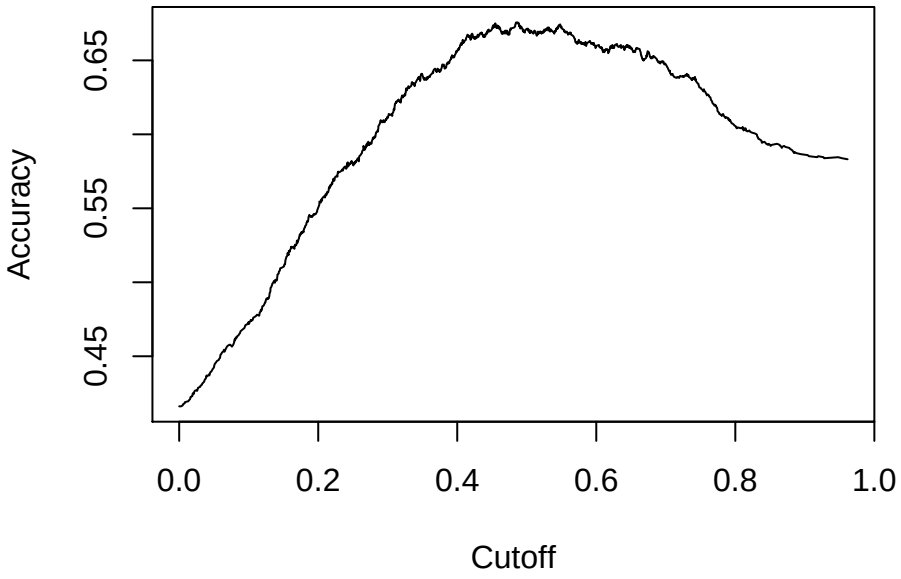
My mentor suggested that I could treat the problem as a classification problem, whether deaths occurred or not. So I divided the data set into two groups, one with deaths and one without fatality, and saved this as a new variable. Then I used this newly created variable and ran the random forest algorithm again. The out-of-bag error is 32.85%. We can take a look at how well our model is by plotting the ROC (Receiver

Operating Characteristic) curve. Here we see that the area under curve (AUC) is 0.72, which means that there is 72% chance that our model will rank a randomly chosen postive example (deaths exist) higher than a randomly chosen negative example (no deaths).

```
##
## Call:
##  randomForest(formula = DEATHS_EXIST ~ EQ_PRIMARY + FOCAL_DEPTH +      LATITUDE + LONGITUDE + Popula
##                Type of random forest: classification
##                      Number of trees: 1000
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 32.85%
## Confusion matrix:
##     0   1 class.error
## 0 640 199   0.2371871
## 1 273 325   0.4565217
```

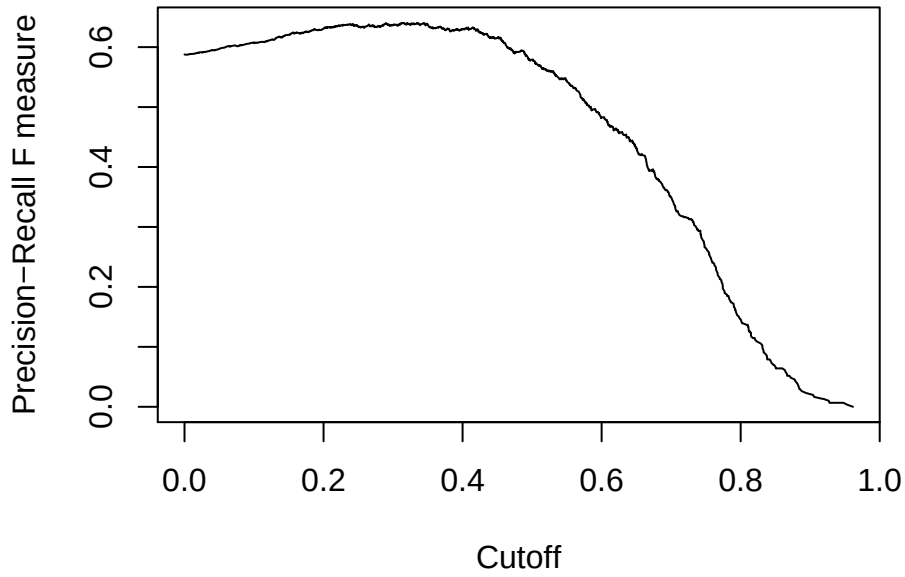


I plotted the accuracy plot to show how the accuracy changes as a function of cutoff value. And we can see when the cut off value is ~ 0.5, we achieve the highest accuracy.




```
##  accuracy    cutoff
## 0.6757133 0.4869565
```

Similarly, we can show how the f-score changes as a function of cutoff.



Using this classification model, we can try to make predictions on the fatality at a specific location, given the earthquake magnitude, depth, the population and GDP. For instance, if a magnitude 7 earthquake at a depth of 32.97 km hits a country with 256 million people and 917 billion US dollar GDP, the prediction on whether there will be fatalities is shown below. The plot looks rather blocky, likely because of the interpolation of data at areas without data points. In addition, as we noted earlier, we have kept the earthquake magnitude, depth, country's population and GDP as constants, and we are only showing how the fatality changes as a function of location. If we take the same modeling strategy and apply it to a small region with local earthquake,

population and GDP data, we will achieve more detailed predictions.

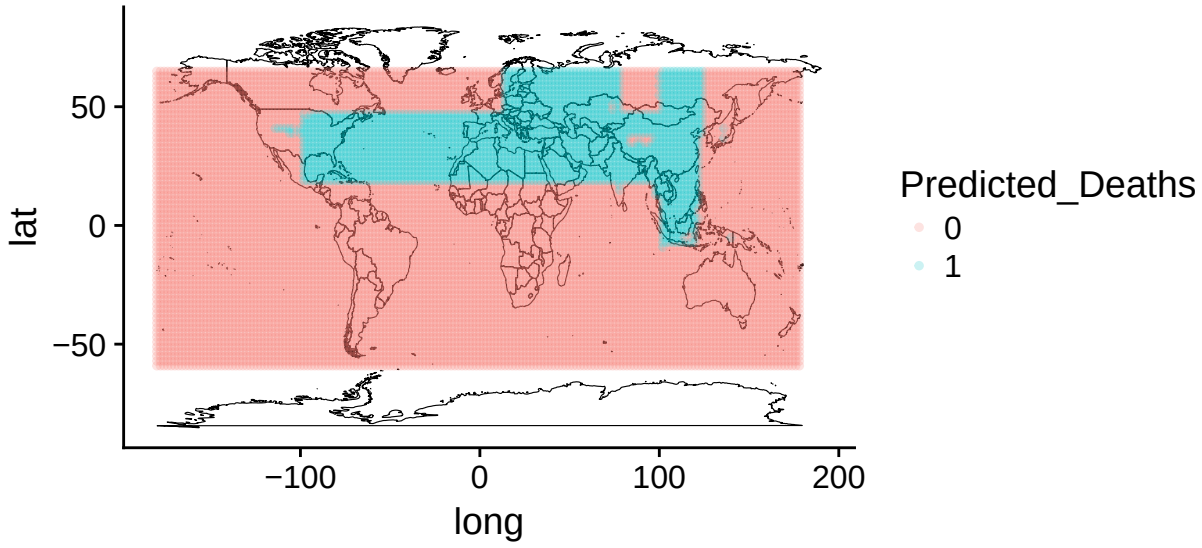
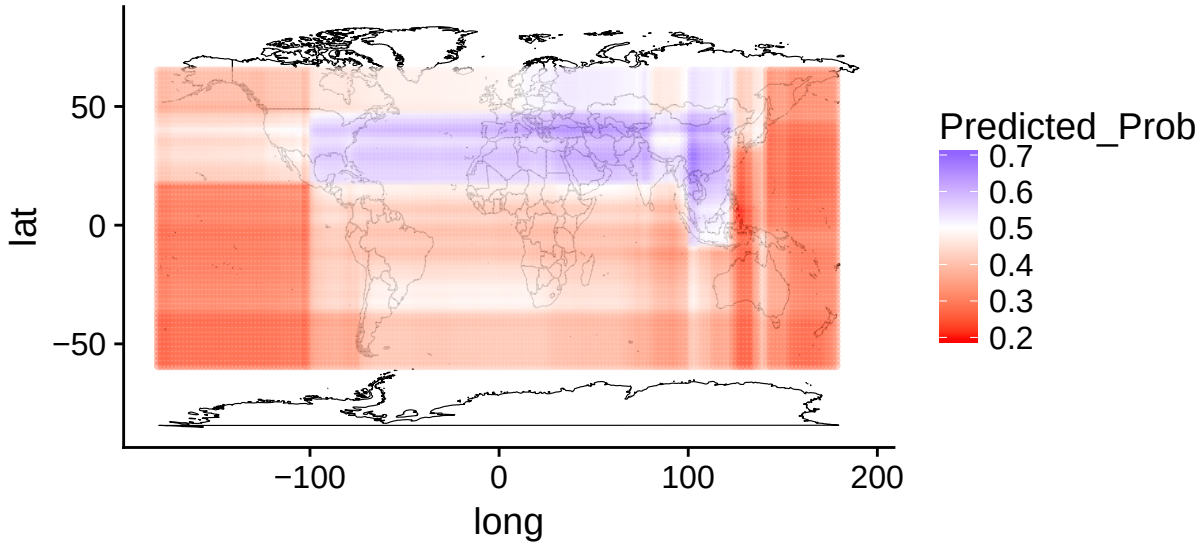
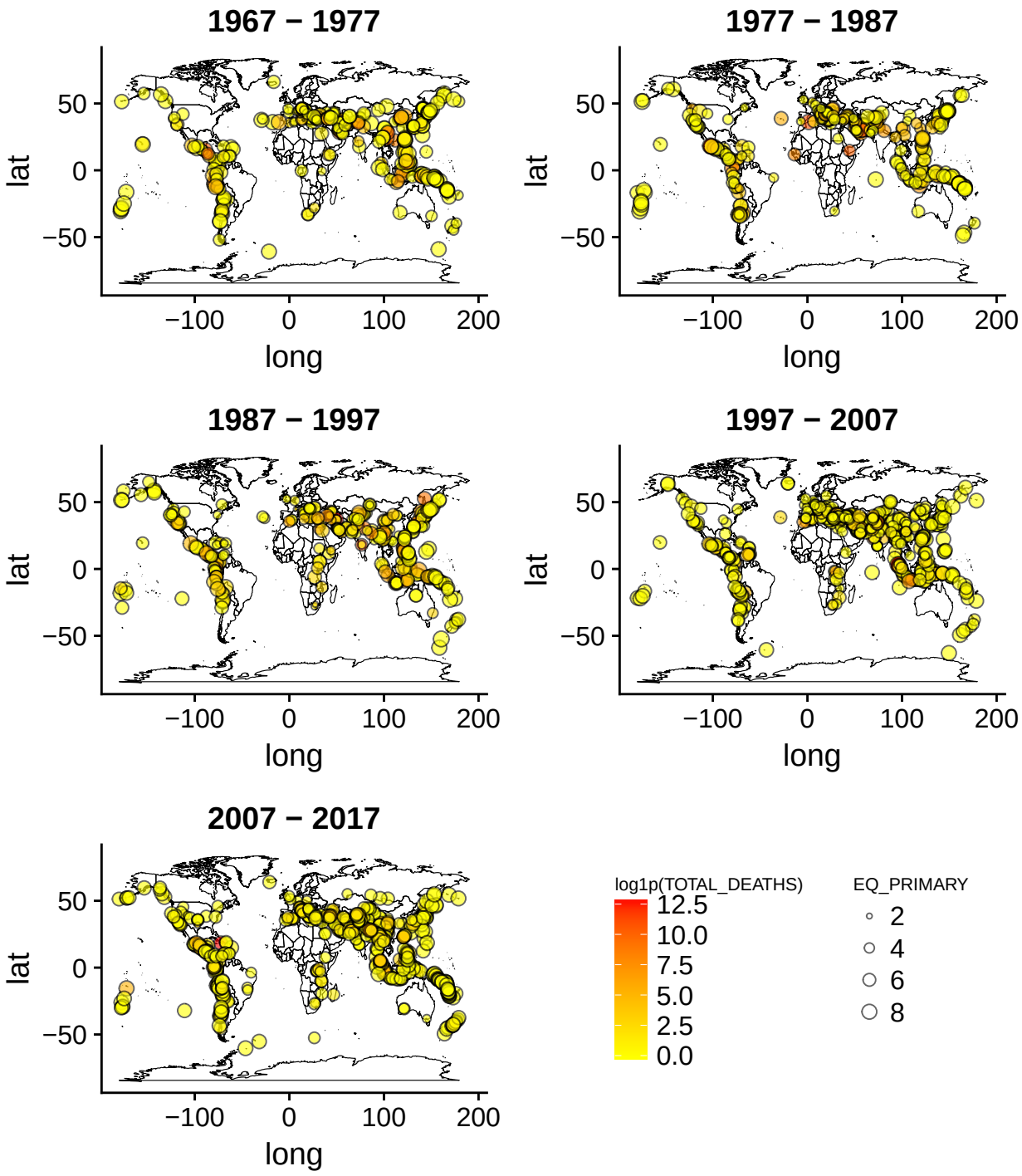


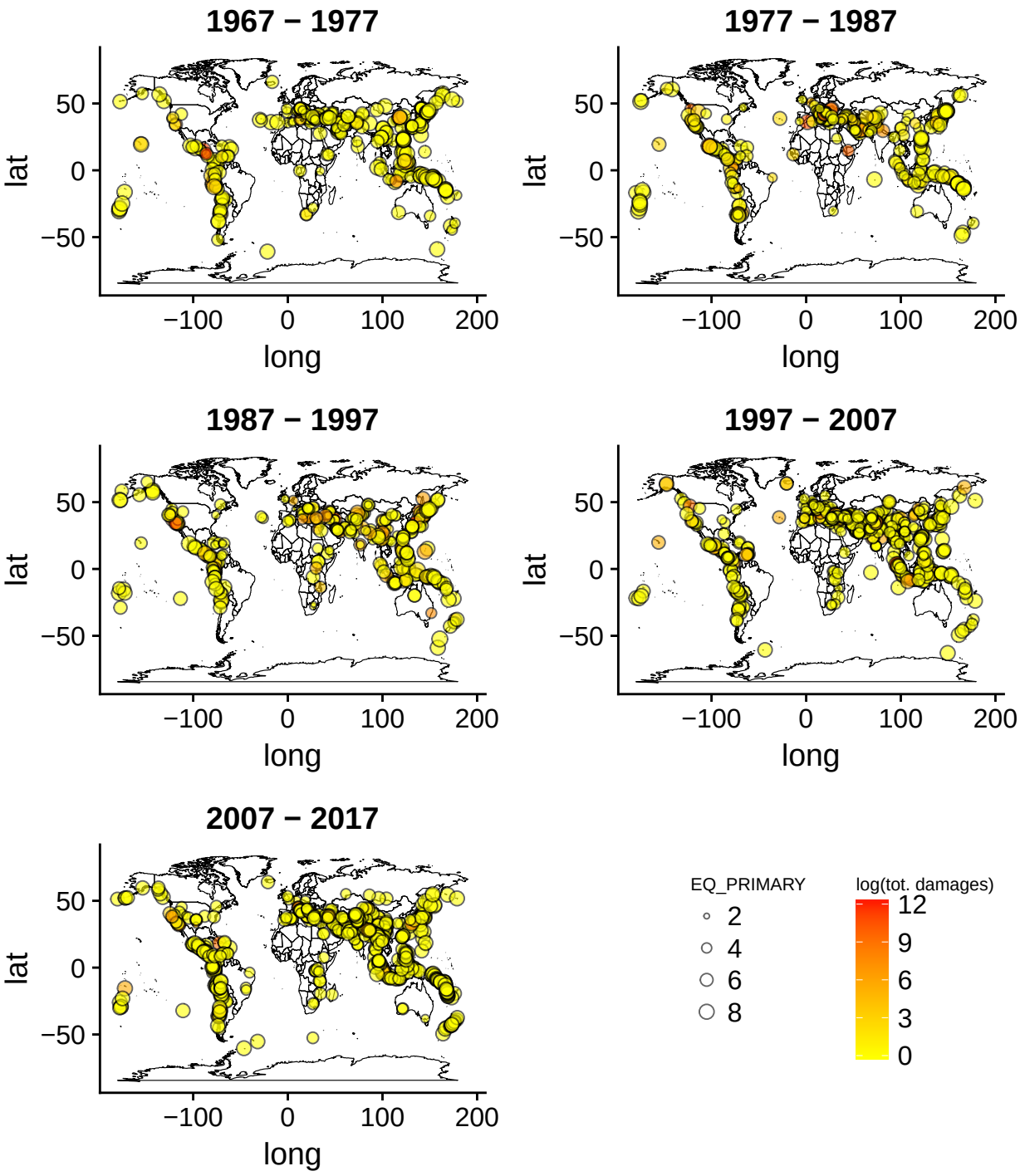
Figure above shows if deaths will occur or not. We can also show the probability of whether there will be fatality. In the figure below, the darker blue area represents greater chance of deaths, and the darker red marks minimal probability of deaths. The pattern looks similar to the figure agove, but it gives information on how likely deaths will occur based on my model.



Map the Earthquakes

While it is useful to explore the correlation between total loss in large earthquakes and the earthquake characteristics, the public would be more interested in learning about where are these earthquakes and how the total loss changes at each location over the years. Using the "ggmap" package in R, we can explore the spatial and temporal change of total loss in large earthquakes. As shown in the following two figures, earthquakes are concentrated along certain areas, which mark the plate boundaries where tectonic plates interact with each other. The data was plotted every 10 years, and by tracking the color of the circles, we can see that for certain regions such as in the Mediterranean and the west coast of South America, the total loss has decreased over the years.


1967 − 1977
1977 − 1987
1987 − 1997
1997 − 2007
2007 − 2017
lat
long
50
0
−50
−100
0
100
200
log1p(TOTAL_DEATHS)
12.5
10.0
7.5
5.0
2.5
0.0
EQ_PRIMARY
2
4
6
8



Summary

By analyzing the significant earthquake data for the past 50 years, I am able to answer a few questions that I had at the beginning. First, overall, the total loss in large earthquakes did not decrease over the years, but the loss seemed to decrease at certain regions, such as the Mediterranean and part of South America. Second, there is a correlation between the total loss and earthquake characteristics. Specifically, large earthquakes and shallow earthquakes are correlated with more damages and deaths. These correlation, however, are not linear, as evidenced by the poor results from the linear regression analysis.

A country's population and GDP are also likely correlated with loss in earthquakes, which were explored in this exercise too. While the linear regression analysis did not reveal interpretable correlation between the population/GDP and the loss in earthquakes. Treating the total deaths as a classification problem, we can explore the deaths as a function of earthquake variables and a country's GDP and population. The model could fit 68% of our data, which is not great but can provide some information on if deaths will occur in an earthquake at a location. Using this model to predict the fatality at a location, given earthquake parameters and GDP/population, maps out where there might be deaths at earthquakes. Due to the lack of information on fault distribution and variation of population/GDP for each area, interpretation of the prediction results calls for caution though.

Visualization of the data seems to reveal that rich countries (countries with higher GDP) are subject to less loss in earthquakes. This correlation requires further investigation too because the distribution of earthquakes and GDP are not uniform across a country. Plotting the total loss in an earthquake which only affect a small region over an entire country's GDP may have over simplied the problem. A clearer correlation may emerge if we consider regional GDP, instead of the national GDP.

Results from this project will be useful for the UN to budget emergency fund for earthquake-prone areas, especially when combined with earthquake risk analysis results. For areas that are suffering great loss over the past 50 years and still have high risk of earthquakes, a greater amount of emergency fund should be budgeted. The analyses performed in this project can be furthered to study similar correlations in individual country, state or even city. And the results can be used to identify countries, states and cities that have succeeded in reducing the total loss over the years. Strategies deployed by these governments can be borrowed by places where earthquakes related damages/deaths are still very high.