Capstone Project Report

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

Chen Chen

12/10/2017

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 data, 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.

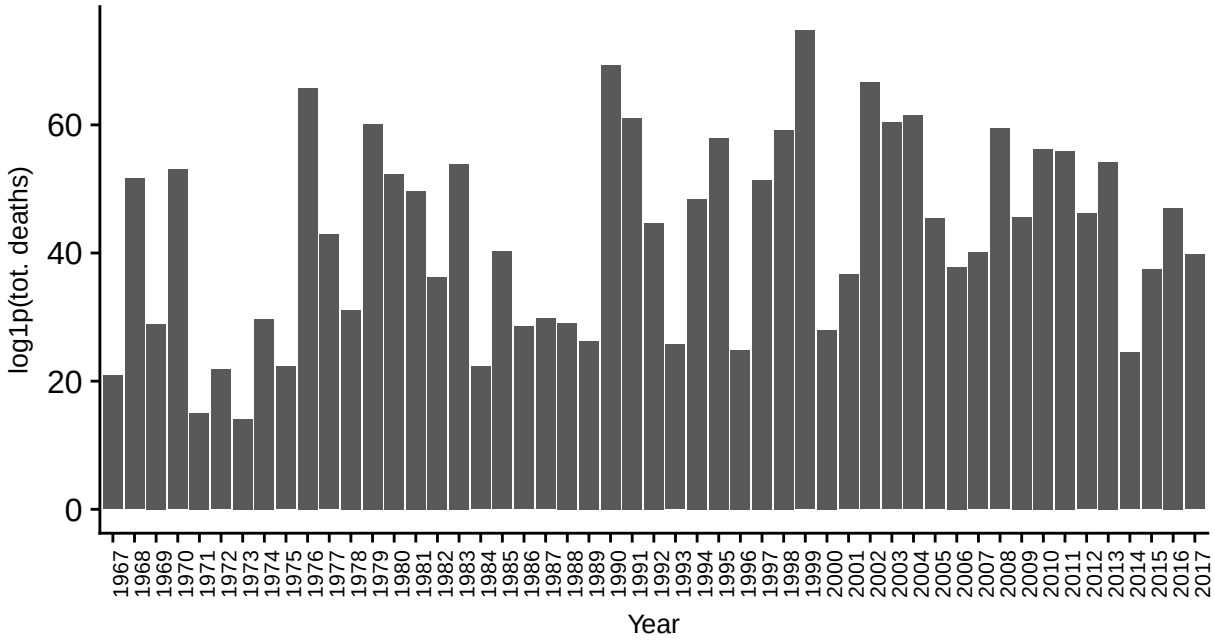
One limitation of this dataset is the absence the local population density information at each earthquake location. While a country's population is an indication of how dense the population is, the distribution of population varies from area to area. In addition, large countries will have great values of population, but the earthquakes may occur along faults that are far away from the residential areas. Another limiation is the lack of information on building code/style at each earthquake location. One would image 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.

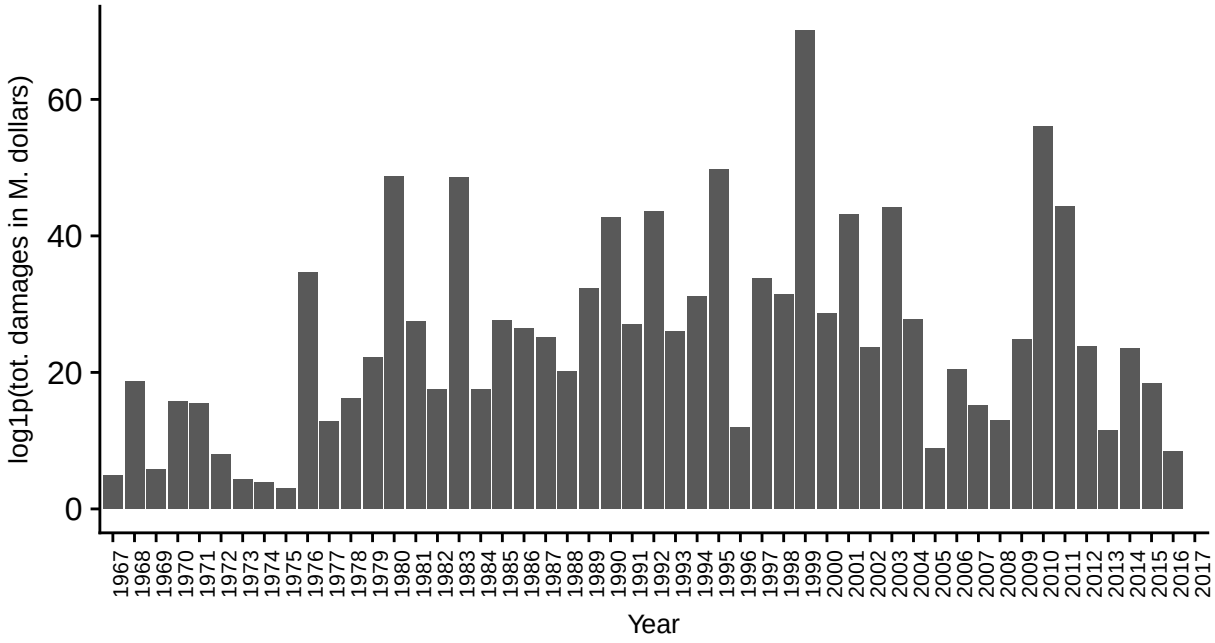
To clean up the data, I converted the date and time of the earthquakes from characters to date in R. Using the "dplyr" and "tidyr" packages, I re-arranged the earthquake data by countries and earthquake variables that were used in the analysis. I also collapsed the GDP and population data into Year-Value 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

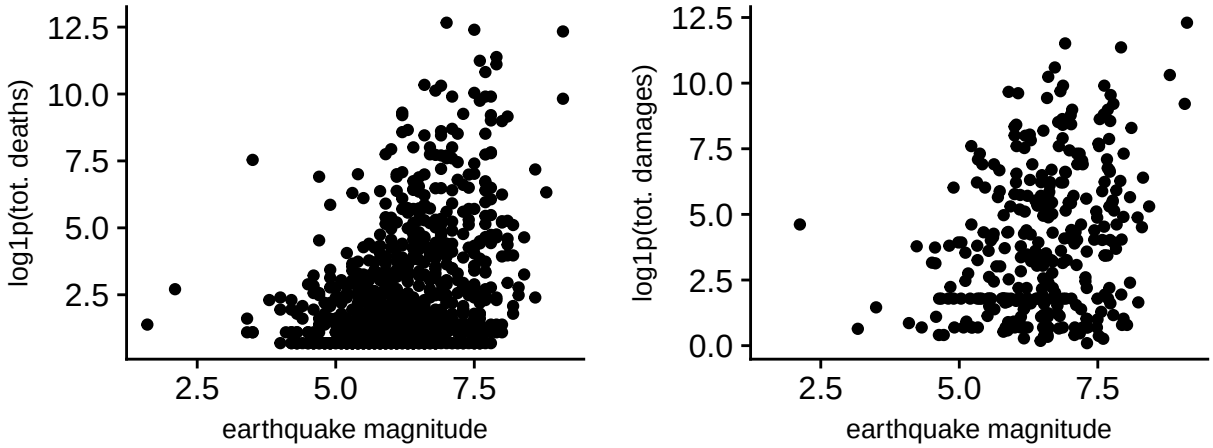
Visualization of the Data

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.

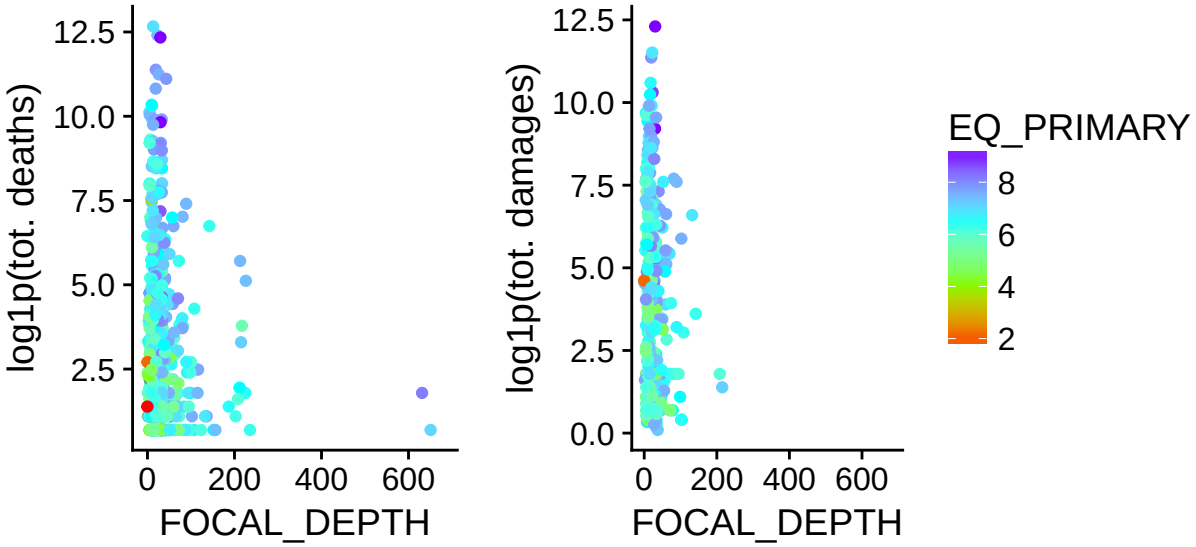




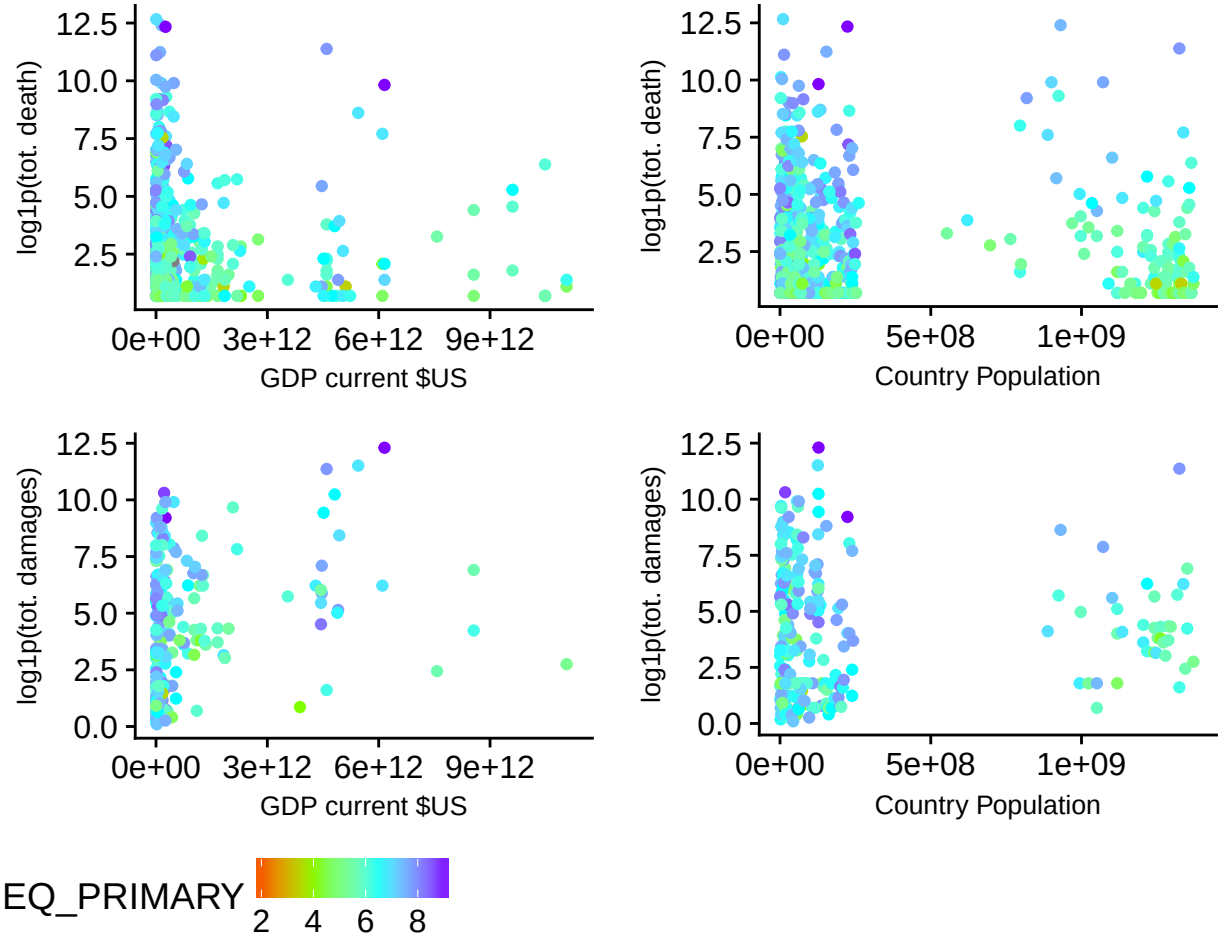
There is a positive correlation of earthquake magnitude and the total death/damages in an earthquake. As expected, more people died and more damages occurred in large earthquakes. Instead of a linear correlation, the distribution of the data points is more spreaded out.



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 had 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 predictor (total deaths/damages) is 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
## -12574  -5071  -2347    822 229070
##
```

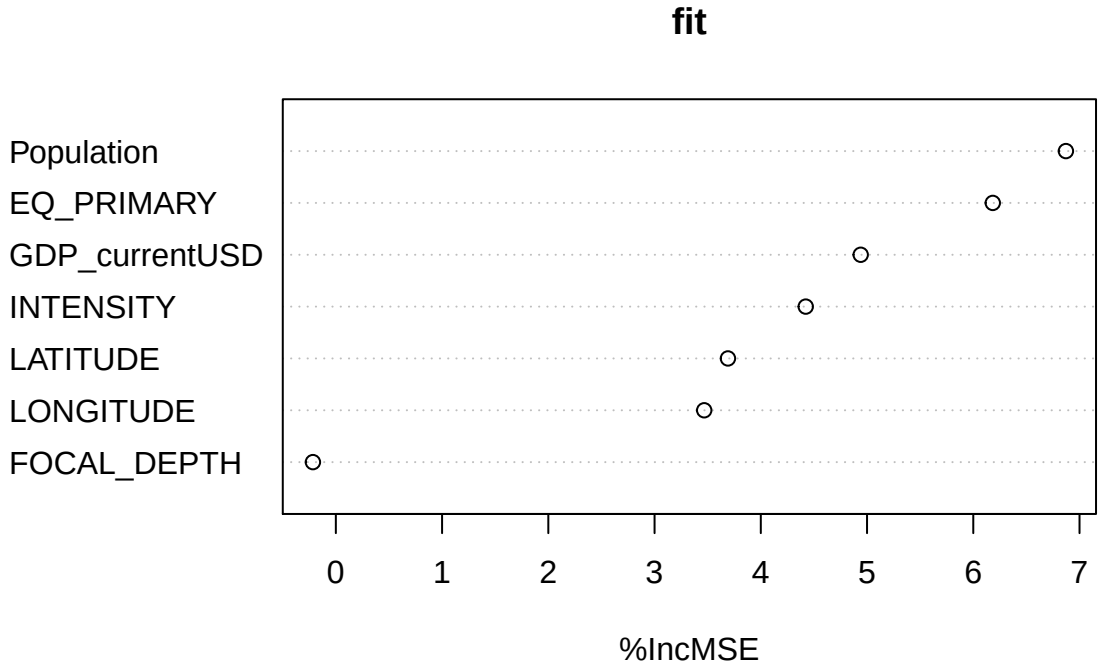
```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -29396.051   9095.856  -3.232  0.00139 **
## FOCAL_DEPTH     -4.789     38.116  -0.126  0.90011
## EQ_PRIMARY    2579.113   1381.193   1.867  0.06302 .
## INTENSITY     1872.967    786.637   2.381  0.01801 *
## LATITUDE        50.497     52.875   0.955  0.34049
## LONGITUDE       10.689     13.159   0.812  0.41737
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17410 on 251 degrees of freedom
##   (1755 observations deleted due to missingness)
## Multiple R-squared:  0.06531,    Adjusted R-squared:  0.04669
## F-statistic: 3.508 on 5 and 251 DF,  p-value: 0.004401
```

I also included the economic status of a country represented by the GDP and the population in the regression analysis. Inspecting the summary of this model suggests a slight increase in performance of this model, 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
## -22102  -5239  -1374   1977 219126
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -3.603e+04  1.157e+04  -3.114  0.00214 **
## FOCAL_DEPTH     2.414e+01  6.273e+01   0.385  0.70083
## EQ_PRIMARY      3.053e+03  1.727e+03   1.768  0.07867 .
## INTENSITY       2.047e+03  1.006e+03   2.035  0.04330 *
## LATITUDE        3.940e+01  6.644e+01   0.593  0.55390
## LONGITUDE      -9.284e+00  1.778e+01  -0.522  0.60217
## GDP_currentUSD  2.611e-10  1.753e-09   0.149  0.88177
## Population      1.418e-05  4.637e-06   3.058  0.00256 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19070 on 186 degrees of freedom
##   (1818 observations deleted due to missingness)
## Multiple R-squared:  0.1173, Adjusted R-squared:  0.08412
## F-statistic: 3.532 on 7 and 186 DF,  p-value: 0.001382
```

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

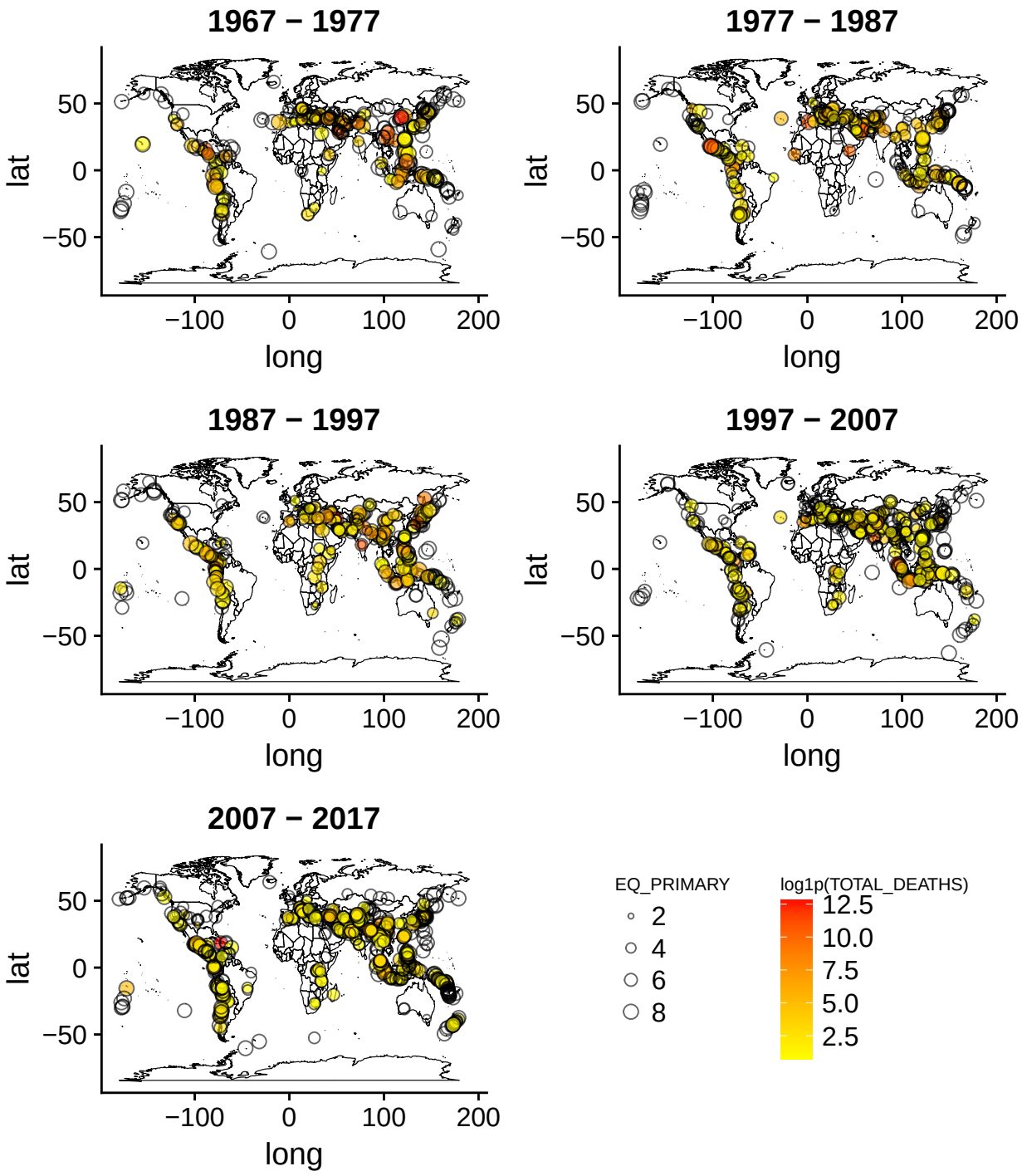
adequate to define a good fit model that relates the total deaths/damages to the earthquake variables.

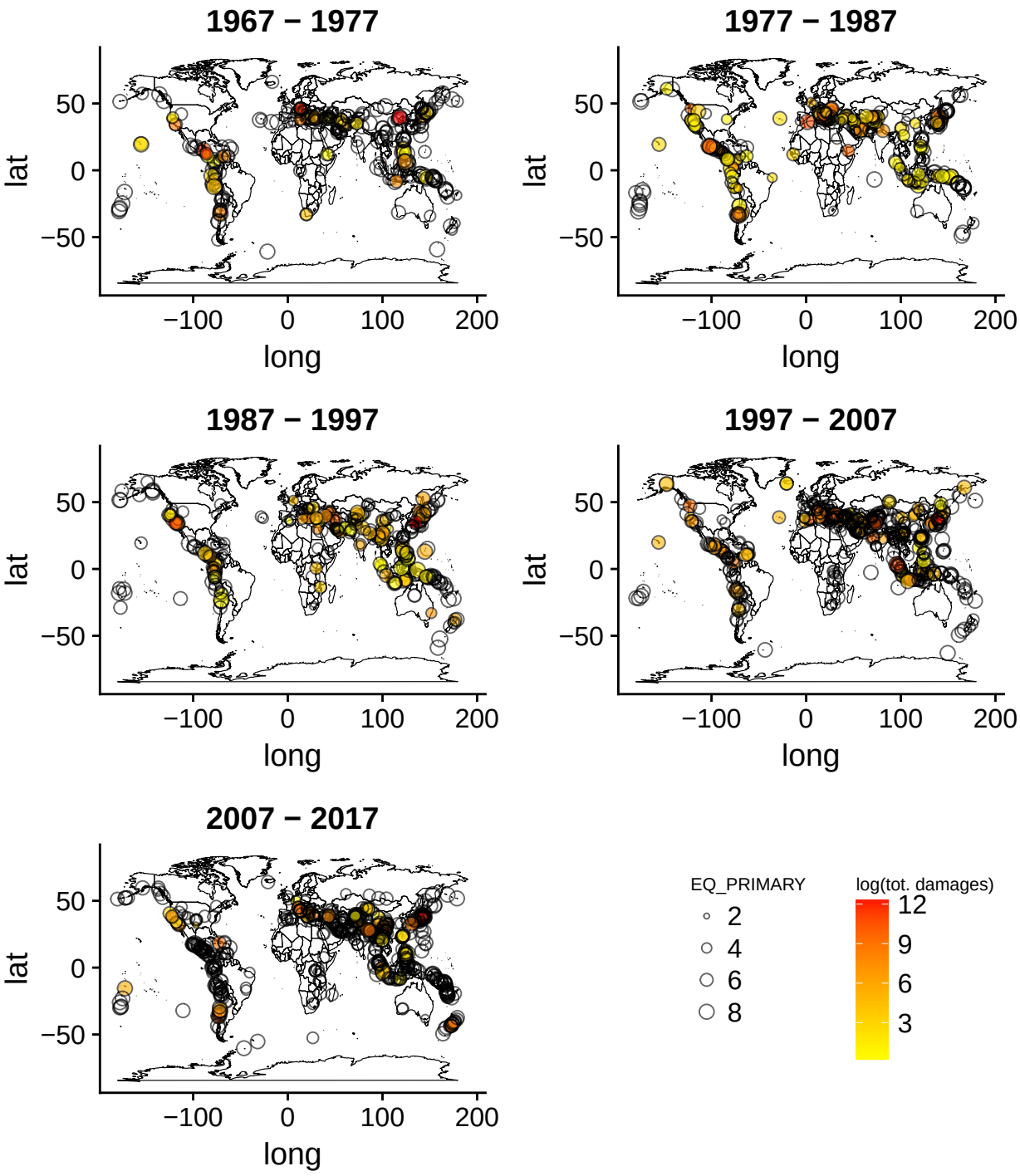



```
##
## Call:
##  randomForest(formula = TOTAL_DEATHS ~ FOCAL_DEPTH + EQ_PRIMARY +      INTENSITY + LATITUDE + LONGITU
##                Type of random forest: regression
##                      Number of trees: 2000
## No. of variables tried at each split: 2
##
##           Mean of squared residuals: 408750305
##                     % Var explained: -3.45
```

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
EQ_PRIMARY
2
4
6
8
log1p(TOTAL_DEATHS)
12.5
10.0
7.5
5.0
2.5



Summary

By analyzing the significant earthquake data for the past 50 years, I am able to answer a few questions that I had. First, the overall 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 and the random forest method did not reveal interpretable correlation between the population/GDP and the loss in earthquakes, 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 because the disbution of earthquakes and GDP are not uniform across a country. Plotting the total loss in earthquakes which only struck 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 for 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.