

Good Vibes: The Probability of Earthquake Magnitude and Death

A Proposal for the Introduction to Data Science Project

Team Fugacity Members

Paul Mundt, u0932146

Katie Jones, u0541901

Motivations

Last year on February 15th, 2019 at 05:09am (MST) a 3.7M earthquake hit Bluffdale, UT. The magnitude was enough for residents in neighboring cities such as Draper and Riverton, to feel the ground shake. This caused many people in the area to wonder about their safety as well as their preparedness for if another earthquake were to occur. As residents of a city built on a faultline, we are keenly aware of the ubiquity of earthquakes to life in Salt Lake City.

Earthquakes happen all over the world, most of which we cannot feel. As citizens living along a major faultline, the question of earthquake magnitude and danger is always a concern. This project will look at the historical data of earthquakes around the globe and determine the danger of displacement, property damage, injury, and death (called impact factors from this point forward) in relation to earthquake magnitude and other socioeconomic factors.

Project Objectives

For this project, we are trying to answer the following questions for a specific set of locations:

- 1) What are the magnitudes of the earthquakes within a 50-mile radius of the location?
- 2) How often do earthquakes at each location occur?
- 3) What is the probability of earthquakes at each magnitude for each location?
- 4) What is the probability of displacement, property damage, injury, or death in an earthquake?
- 5) What other information, like GDP of the country, population density, etc, affects impact factors for a location?

We will be investigating these claims by analyzing seismographic information about earthquakes, information about effects of earthquakes, and looking at a variety of other data types (national GDP, population, population density, etc) . Next, we hope to apply the information gleaned from the data to a geomap for a visualization as well as creating a boundary for the number of earthquakes in an area.

Data and Sources

Data

These data sources contain information about earthquakes and their magnitudes, effects (injury, property damage, mortality, etc.), world population data, world GDP data, world population density, etc. Other data sources may be used as the project is further along. A majority of the data that we will be using will come from the following sources:

- US Geological Survey
- Earthquake Impact Database by Earthquake-Reports
- United Nations- Department of Economics and Social Affairs
- The World Bank
- NASA EarthData and Socioeconomic Data and Applications Center (SEDAC)

Links to Data Sets

Earthquake Locations:

http://feed.unmung.com/feedfeed=http://earthquake.usgs.gov/earthquakes/feed/v1.0/summary/all_week.atom
(http://feed.unmung.com/feedfeed=http://earthquake.usgs.gov/earthquakes/feed/v1.0/summary/all_week.atom)

Past Data: <https://earthquake-report.com/2018/02/10/earthquake-impact-database-2018/> (<https://earthquake-report.com/2018/02/10/earthquake-impact-database-2018/>)

Ethical Considerations

The main stakeholders in a data science project like this are general civilians and people involved in infrastructure (architects, city planners, civil engineers, law makers, etc). To the general citizen, this project can be both incredibly informative as well as fear inducing. Most people only gain exposure to earthquakes after a major disaster. Because of this, many people inherently fear earthquakes and the impact that they have on daily life. For this project, it is important that all correlations and statistics we explore are accurately and accessible explained and given context. As is often the case, the context given to the data is usually more important than what the data itself says.

This project will also be useful for people involved in infrastructure. By showing what variables that affect impact factors, the people in charge of building structures in a city might be better informed on the need for earthquake resistant structures. Areas of frequent high magnitude earthquakes are supposedly safer than those with the semi-active seismic activity. We believe this is because infrastructure and response are more highly valued when building in the area, as more frequent danger incentivises buildings to be designed to withstand more intense oscillation.

Data Processing

We expect some data cleaning to be necessary as many of the data sources have the pertinent information in unuseful data types. For example, the data for coordinates comes as a single line with the letters involved. Those need to be turned into actual float coordinates that correspond to north, south, east, and west. We also need to provide boundaries for countries so we can determine frequency per country. Additionally, many sources of data need to be synthesized into a single useful data frame. This will require that the way different sources catalogue their data will have to be normalized to a useful data structure.

Exploratory Analysis

We plan on using correlation matrix as well as probability assessment to determine what variables play into account for frequency of earthquakes as well as damage and mortality rate. We will also be using past data to build an distinction between disturbance ratings and damage caused from the disturbance. This will help in conditioning our expectations as to what a level 8 disturbance can do as compared to a level 3 disturbance. This will also be important for the fatality and injury in accordance with population density as well.

Analysis Methodologies

We plan to use clustering and regression to explore the data about earthquakes. Clustering can be done to group data by similar characteristics (location, GDP, population, etc.) then regression modeling can be done to develop predictive measures for the danger from earthquakes. We then hope to use the predictive models to see if our predictions match some of the more well known earthquakes from the last decade. Additionally, we want to compute the danger to Salt Lake City for a major earthquake and compare it to expert predictions.

Project Schedule

Sunday, March 8th: Collect complete list of data sources, start scraping

Sunday, March 15th: Clean all data and compile into data frame, begin initial visualization

Sunday, March 22nd: Begin clustering and regression

Sunday, March 29th: Submit Milestone, continue regressions

Sunday, April 5th: Begin geo-mapping interface, complete regressions

Sunday, April 12th: Complete geomapping, complete final report and video

Sunday, April 19th: Submit Project