Data Science Project

Jaxson Fryer, Ellise Putnam, Pablo Chang Huang

Spring 2023

1 Overview

Whats the data set: CSV file with data from the past 50 years seismic events with magnitude equal or higher than 5. The data set contains Time, Latitude, Longitude, Depth, and Magnitude. The data comes from the USGS earthquake catalog [2]. Additionally, we are using a collection of fault line data from the Americas that will further be used to analyze how certain fault lines influence and effect the locations of earthquakes around North, Central, and South America [3].

What is our motivation for choosing it: Earthquakes are currently considered the most destructive natural hazard. As cities grow bigger and taller the consequences of large scale earthquakes can be more fatal if it they are not prevented. Although seismic events are impossible to predict with 100% precision, there are specific zones where tectonic faults causes constant events. After the cold war, the global seismic network joined and grew to track possible nuclear testing. This seismic receivers network make it possible for the scientific community to map with precision and fast every event in the world. Therefore, there is a big opportunity for humanity to use this data set to create models and find patterns in these events. The motivation to manipulate the data and create models is to better understand earthquakes hot spots over North, Central, and South America as well as the fault lines that influence these earthquakes. It is important for humanity to gather data to find patterns on seismic events and

fault line locations that can be used for city planning, hazards management, or humanitarian work as population grow rapidly.

What are we solving with the data set: Through analyzing earthquake data and taking a small slice of fault line data from the Americas we can create a model to predict where a fault line is. We are going to do this by splitting the fault line data and earthquake data into training and testing sets and use the separate sets to improve our algorithm. We are also going to predict and deliniate between different fault lines.

2 Data Acquisition

The data set our team is working with is a CSV file with data from the past 50 years seismic events with magnitude equal or higher than 5 and fault line data of the second half of the ring of fire down the Americas. The data comes from the USGS earthquake catalog [2] and Kaggle [3]. The data sets have a lot of elements in them such as: Time, Longitude, Latitude, Magnitude, Magnitude Error, Depth, Depth Error, What station(s) took the measurements, and much more information that is not useful for our analysis. We will extract the longitude and latitude first to determine where that earthquake took place. Then we will extract the magnitude and the depth to determine how to represent that data point on the Geo-graph. The main limitation of our data is that a of the elements are not useful to do analysis on. The data set is also going to require a lot of preprocessing before it comes usable.

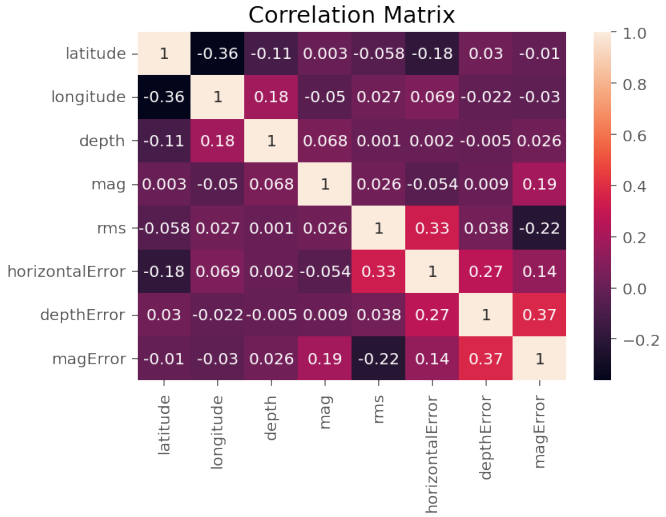
3 Preprocessing

We had a lot to work through for preprocessing. First we had to alter the date format from a serial number (Reference Serial Number: 2022-12-31T03:31:43.830Z) so it could be read into our data frame in the format we wanted (Reference Modified Format: 12/31/2022). Then we removed the elements we deemed unusable or unnecessary for our data frame. For instance we removed a column titled

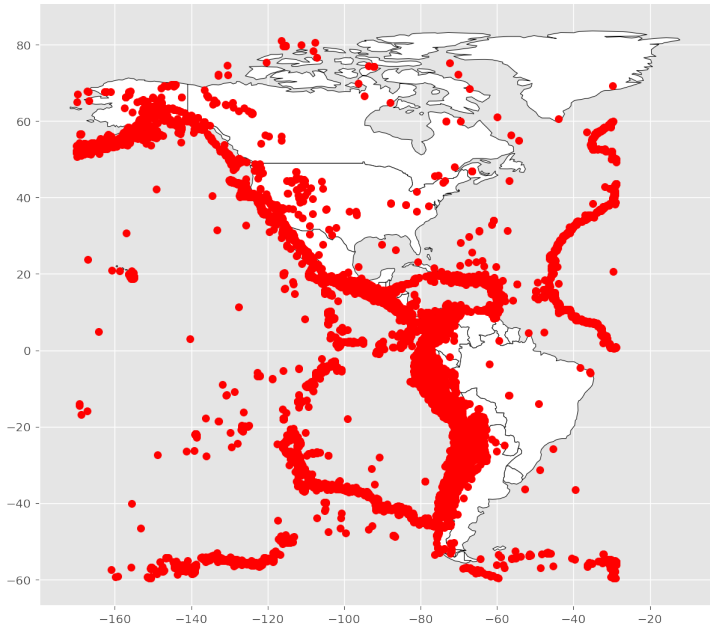
status which only gave us information on whether the earthquake was reviewed by USGS or not. After filtering the csv file with the data contained the columns: Date, Latitude, Longitude, Depth, Magnitude, RMS, Horizontal Error, Depth Error, and Magnitude Error. The data is going to be processed and graphed into maps with the GeoPanda library that has graphing and data manipulations relevant to the project. To start the analysis the raw data of every earthquake was mapped in the Americas representing with a red dot the location of every event. The map is shown below and it is the start of making connections between the earthquakes and the locations:

We wanted to create a couple visualizations to better understand the data.

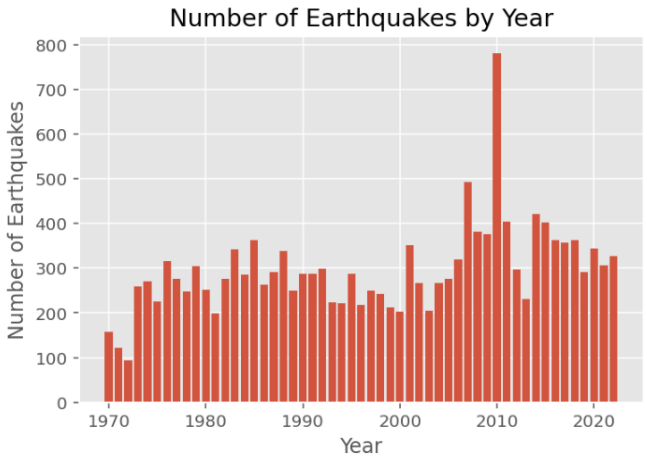
First we made a correlation heat map of all of the variables before doing any data cleansing and found that very few of our data correlated at all. So we then generated a correlation heat map of the data that will be most usable for our analysis.



Second we created a Geo-graph of North and South America to create a heat map of what regions have the highest frequency of earthquakes.



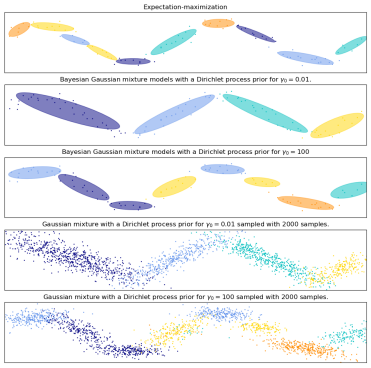
Third we created a bar chart of the amount of earthquakes from our 50 year period to see if there was a cyclical pattern in earthquakes over the years. Our goal with this bar chart was to hopefully find a pattern and from that be able to try and predict 5 year time frames with the highest chance of having more earthquakes. Based on the bar chart we cannot make a clear conclusion about a consistent cyclical pattern, however, something that we could take into consideration for the future might be looking at different regions and performing the same process in hopes of finding a pattern and predicting 5 year time frames with higher chances of more earthquakes in that region.



4 Model Selection

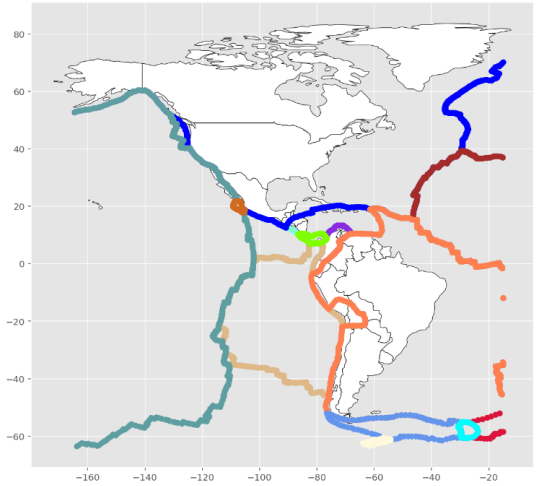
We picked Gaussian Mixture as our prediction clustering model to predict plate boundaries locations or region division with earthquake longitudes and latitudes. Gaussian Mixture is what we believe is the ideal algorithm for our problem because it is really good at finding "clean" lines in extremely noisy data. The model also helps with uncertainty about the correct number of clusters which is extremely beneficial because we have so many clusters. The model works through a probabilistic model that assumes all the data points are generated from a mixture of a finite number of Gaussian distribution. The most important part of this model is that the class can adapt its number of mixture components automatically. On this way we can determined the division of different earthquakes per plate boundaries by changing the sensitivity of the model. There are different types of Gaussian mixtures, there are Bayesian mixture, Gaussian mixture with Dirichlet process, expectation maximization, etc. All of the the specific used model will be defined depending on the model testing, and training results. All of the models work similarly in diving data by regions based on the data correlations, creating sub populations within the data. An example of the application of the different models can be found above showing how it can be

used to get the data sets of earthquakes divided by different plate boundaries. Below is an image example of how the Gaussian Mixture algorithm works.

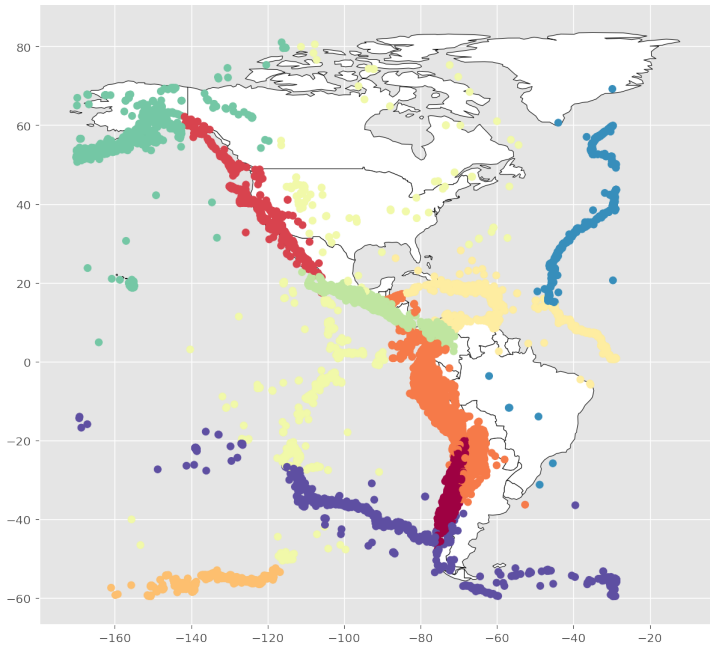


5 Results and Evaluation

Using our Gaussian Mixture model to evaluate plates and plate boundaries we were able to predict with some accuracy plates and plate boundaries. As you can see in the below picture these are the actual plate locations and plate boundaries. One of the techniques we had to learn was how to overlay two plots of different types (scatter and normal plot) we were able to solve this by aligning the axis and then overlaying the plots.



By looking at the graph above with the plate locations there are 13 defined plates. However, while we were able to define 13 different plates, when looking at the graph below, we were only able to depict 12 clusters mainly along the western coasts of North, South and Central America. However we were able to delineate plate boundaries very clearly as you can see by the below graph. Colors represent different plates and the individual dots represent boundaries.



While we don't have a quantitative metric that we can analyze, we can compare the two graphs above in relation to how the plate boundary locations were able to be predicted as accurately as they were. The first graph, as previously discussed, locates the plate boundaries. In the second graph, while only identifying three colored clusters, we are able to see that the pattern and location of the plates aligns with the pattern and locations of the actual plate boundaries in the first graph.

The reason we did not come up with more quantitative metrics for our data was because it is very difficult to compare plate boundaries to predicted plate boundaries to predict the accuracy because we would have to use a clustering

algorithm on the testing data points and compare them to a clustering algorithm of the predicted data points. This would make our testing metric very inaccurate because we would be estimating our predictions and estimating our testing. That's why it is not as simple as the projects we did in class because you are not comparing test point A to predicted point A.

6 Sources

1. https://www.usgs.gov/programs/earthquake-hazards
2. https://earthquake.usgs.gov/earthquakes/search
3. https://www.kaggle.com/datasets/cwthompson/tectonic-plate-boundaries?resource=download
4. https://scikitlearn.org/stable/auto_examples/mixture/plot_gmm_sin.html#sphx-glr-auto-examples-mixture-plot-gmm-sin-py