CS470 Final Project

Leonardo de Farias, Brian Weinshenker, James Broderick, Zach Daube

December 2025

Collaboration Statement

Leonardo, Brian, James, and Zach acknowledge that we collaborated and worked in parallel to complete this project. We did not use outside help from a tutor, A.I. generated code, or otherwise.

Resources Used:

- Github
- Jupyter notebook
- StackOverflow
- GeeksForGeeks
- MatPlotLib
- sklearn
- Additional sources for research and database references can be found in Works Cited.

Problem Description

Background/Motivation

The majority of detected earthquakes happen along or extremely close to well-defined tectonic plate boundaries and fault lines, where seismic activity is predictable and generally expected. However, on some occasions, earthquakes occur in regions far from tectonic plate boundaries that would otherwise be considered geologically stable. These anomalies can be caused by a multitude of factors, some natural, such as ancient fault lines and stress transfer from other earthquakes, and some artificial, from human activity like fracking and mining operations. There are several reasons for identifying and characterizing these events:

- Major fault lines, or areas of increased seismic activity, can be inferred before being discovered by sonar, sensors, satellite imaging, or other scientific processes.
- They may provide insight into potential future seismic regions or into areas with emerging seismic hazards.
- They could reveal the effects of human artificial activities on seismic stability.

Problem Statement

This project aims to address the challenge of automatically distinguishing between earthquakes that occur in generally accepted zones of seismic activity and anomalous earthquake events that occur in stable regions using unsupervised learning. We specifically will:

- Apply DBSCAN (Density-Based Spatial Clustering of Applications with Noise) to the Global Earthquake Dataset from Kaggle to identify clusters corresponding to known fault lines
- Find earthquakes considered outliers (noise points) from DBSCAN, and then analyze their spatial distribution and maginitdue patterns as well as other characteristics
- Analyze the clusters returned by DBSCAN and investigate the distribution of earthquakes within clusters to identify any patterns or insights

Data Description

Dataset Source

This project utilizes the "Significant Earthquakes, 1965-2016" dataset compiled by the National Earthquake Information Center (NEIC) and found through Kaggle. The dataset contains comprehensive records of all earthquakes with a reported magnitude of 5.5 or higher from 1965 to 2016.

Dataset URL: `https://www.kaggle.com/datasets/usgs/earthquake-database`

Dataset Scope and Size

The dataset has approximately 23,400 earthquake records from 1965 to 2016, spanning five decades. Each event has detailed information on its date, time, location, depth, magnitude, and seismological measurements. The choice to only include earthquakes of magnitude ≥ 5.5 and above means the dataset captures what we consider all significant seismic events while ignoring/filtering out minor events. This dataset provides a comprehensive view of seismic activity across the globe in recent decades.

Important Attributes

- **Date and Time**: Timestamp of when the earthquake occurred
- **Latitude and Longitude**: Geographic coordinates (in decimal degrees) specifying the location of the earthquake. We use these as the primary features for our clustering.
- **Type**: Classification of the event
- **Depth**: Depth of earthquake's hypocenter below the surface, measured in kilometers
- **Magnitude**: Earthquake strength, ranging from 5.5 to 9.1 on the Richter scale
- **Magnitude Type**: The scale used to measure magnitude, which can be:
  - ML: Local (Richter) magnitude
  - MS: Surface wave magnitude scale
  - MB (Mb): Body wave magnitude scale
  - MW (Mw): Moment magnitude scale
  - MD (Md): Duration magnitude/signal duration
- **ID**: Unique identifier for each earthquake event.
- **Source**: Network or organization that originally reported the earthquake.

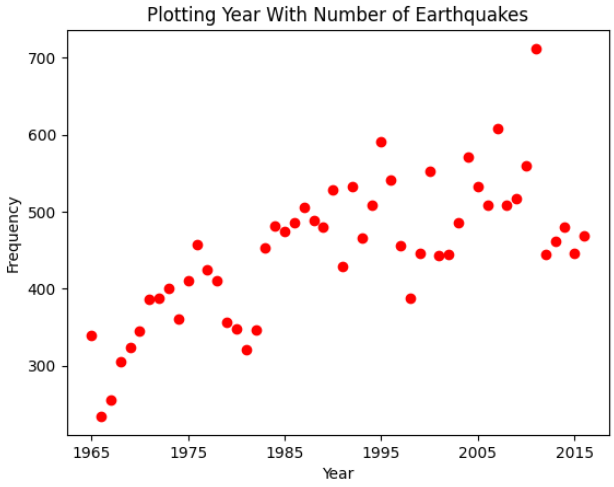


Figure 1: Year vs. Number of Earthquakes

As shown by the plot of year versus number of earthquakes, there is a clear trend of an increase in earthquakes recorded each year, from 1965 to 2016. This is likely due to an increased emphasis on detecting earthquakes by humanity, as well as the advancement of technologies to detect and measure seismic activity.

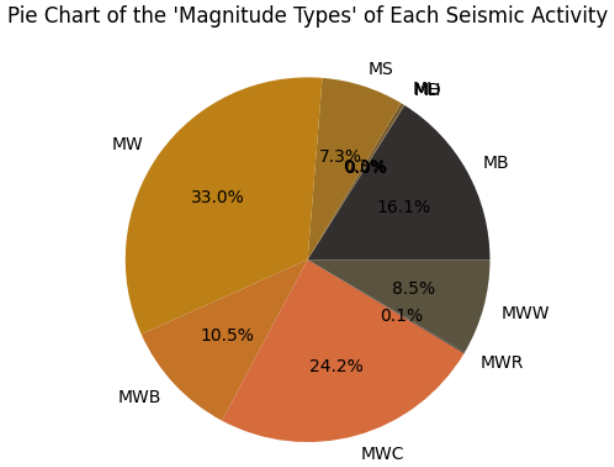


Figure 2: Pie Chart of Magnitude Types in Dataset

There is a good variety in the magnitude types that are represented in sizable proportions, which could be useful for further analysis.

Data Pre-processing

Observation Removal

- **Type**: Vast class imbalance, and only 'Earthquake' class is relevant to our project. Entries of seismic activity type 'Nuclear Explosion', 'Explosion', and 'Rock Burst' were removed.
- **Date/Time**: There were 3 entries with Date/Time in a different format. Since it is such a small portion of the data we simply removed these entries.

Feature Selection (Removal)

- **Time**: Not relevant for our clustering and outlier detection.
- **Type**: Only Earthquake was kept, so the feature is a constant.
- **ID**: Completely irrelevant in earthquake analysis.
- **Source**: We are assuming the sources are reliable.
- **Depth Error, Depth Seismic Stations, Magnitude Error, Magnitude Seismic Stations, Azimuthal Gap, Horizontal Distance, Horizontal Error, Location Source, Magnitude Source, Status**: Significant portion of values are missing and/or feature is not relevant to our problem.

Data Mining Methods

DBScan Algorithm

We implemented the DBScan algorithm in Python. The main points are as follows.

- **Distance Metric**: We used Haversine distance as our distance metric between points. Haversine distance uses latitude and longitude to calculate distance between points on the globe. Using this metric instead of something like Euclidean distance is necessary because the surface of the globe is not an XY-plane but a sphere, and points with similar longitude near the equator are much further than points with the same longitudinal difference near the poles. Additionally, points near -180° and 180° longitude, for example, must be considered as neighbors.

- **Algorithm Outline**: The algorithm converts the latitude and longitude of each point into radians and creates numpy arrays to track which points have been expanded, which points are 'core' points (have $<$ minSamples points in their eps-radius), and a numpy array for cluster labels. We then use the sklearn BallTree module to identify the eps-neighbors of each point. This is the one point of our implementation which was not done from scratch. This is necessary because using a traditional kd-tree fails for polar data, and implementing a ball-tree that spatially encodes spherical data is a complex and tedious task, and doing so in Python would necessarily be significantly more complex memory and computation wise than sklearn's C-based implementation, as well as being somewhat outside the scope of our project.

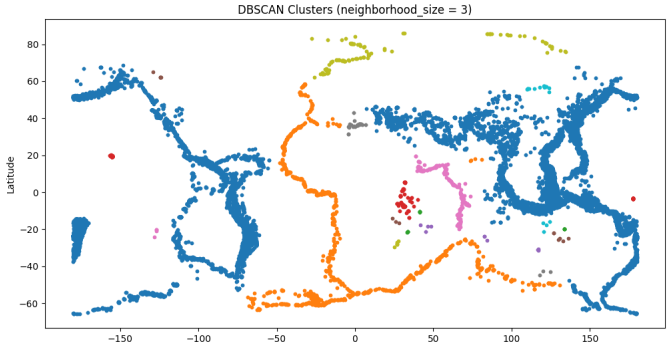
  With the above structures in place, our DBScan algorithm follows the typical operations. We choose a core point to expand and create a cluster starting with its eps-neighbors, continuing to expand each core point that is thus clustered, until no new core points can be reached. Then, a new core point is chosen to begin a new cluster and the steps repeat. Within this process, we consistently used numpy arrays to avoid parsing Python list structures and improve runtime.

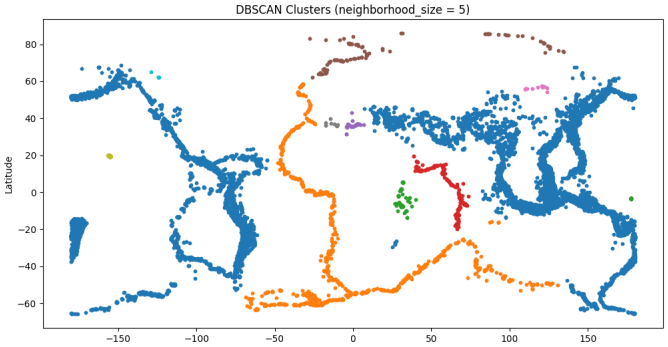
- **Graph Generation**: We pulled publicly available data from the GEM Active Faults Databse and Natural Earth to graph active fault lines and the world map along with our earthquake data using MatPlotLib.

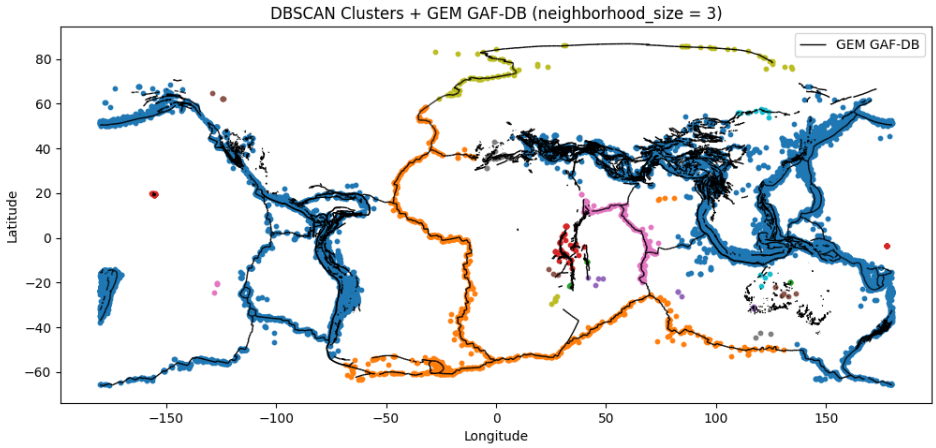
- **Code**: The code for our DBScan implementation can be found in `PROJECT_FOLDER/src/DBScan.py`
  The remainder of our work, including preprocessing and results analysis is modeled in a Jupyter notebook located in `PROJECT_FOLDER/src/main.ipynb`

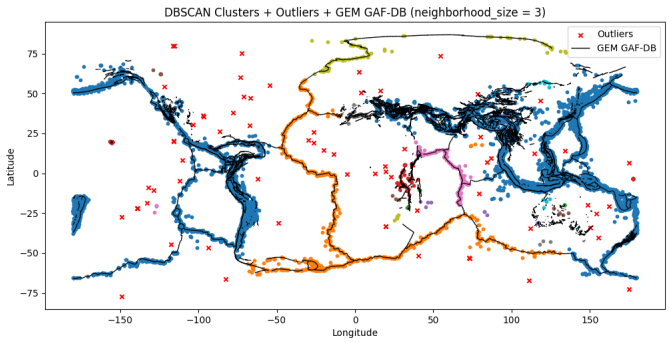
Results

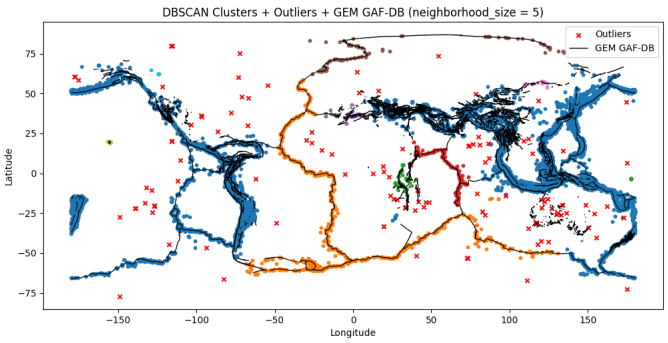
Figures



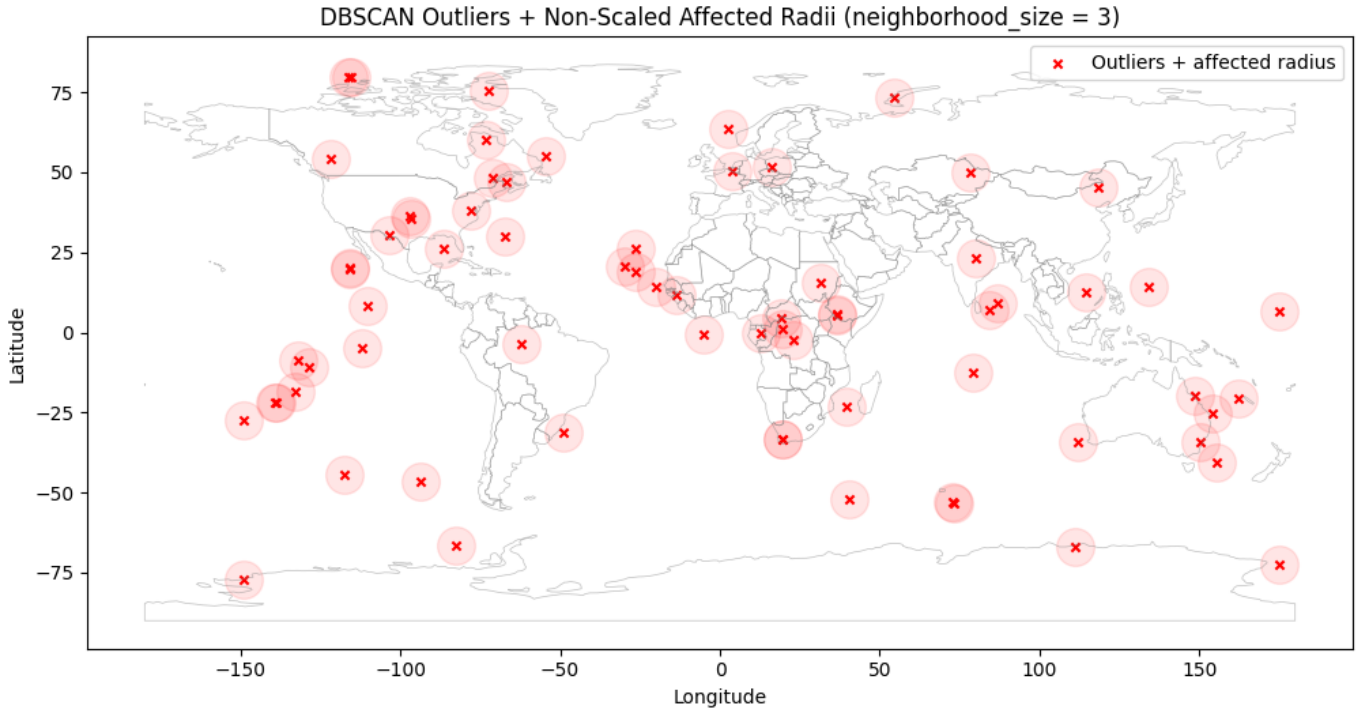






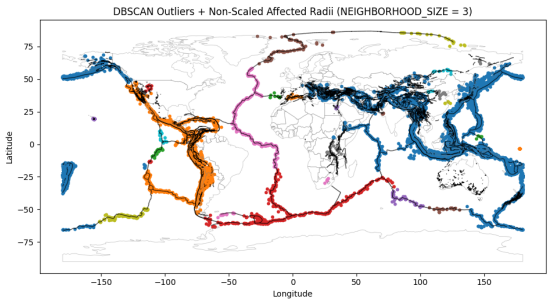


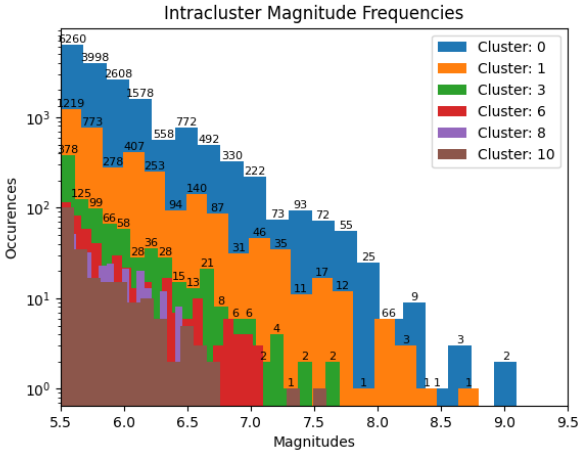
Figures Cont'd



Analysis of Clusters

Below is a frequency table for the magnitudes of resulting clusters and the corresponding cluster map.





$\epsilon = 350$, neighborhood_size $= 5$, cluster_size > 200

Additional intracluster, intercluster, and outlier, statistics (such as country and state of detection using public location APIs) can be found within the Jupyter notebook.

Limitations

Our dataset, although large (23.4k datapoints), only contains seismic events that were recorded by various sensors or satellites around the world. Regions with

sparse monitoring may appear artificially "quiet" in terms of seismic activity. Additionally, the data set only includes seismic events that were recorded as 5.5 or higher in magnitude on the Richter scale. Smaller seismic events, like micro, minor, or light earthquakes do not even appear in this data, and account for the vast majority of global earthquakes.

The DBSCAN clustering algorithm is highly dependent on ϵ (radius) and the min_samples neighborhood size; different values of these parameters may merge distinct faults, improperly split continuous fault systems, or misclassify events as outliers or part of a cluster.

Clustering on latitude and longitude ignores depth, time (sequence of earthquakes), and magnitude trends, which can limit interpretation of fault behavior or temporal evolution (like aftershocks). This model also fails to adequately measure possible shifts in fault patterns or tectonic movement over time; however, the time frame these geological processes typically take to manifest in significant changes in data may as well be infinity, when taking into account the 50 year window this data set encompasses.

Finally, some outliers may be mislocated events, incomplete detection or sensor errors. Not all outliers can be geologically meaningful.

Discussion

Running DBSCAN on the global earthquake dataset ($\geq$5.5 magnitude) produced clusters that closely matched the major fault lines identified by the Global Active Faults Database. This confirms the hypothesis that spatial clustering of seismic events can reveal tectonic plate boundaries and separate, definitive, active fault systems without prior geological information.

Clustering also produced a set of outliers. These represent seismic events occurring outside the dense fault-aligned clusters. Many fall on or near land and exhibit slightly lower average magnitudes than cluster events (see Jupyter statistic comparisons). These outliers may correspond to intraplate seismicity, unmapped fault segments, or areas with limited sensor coverage. This finding is meaningful because accurate identification of both clustered and non-clustered seismic events plays a role in:

- identifying understudied or unexpected seismic zones
- validating or refining global fault maps
- improving risk assessments for regions not traditionally considered seismically active

The fault line vectors were gathered from the public database: Global Active Faults Database, produced by the Global Earthquake Model Foundation$_{(3)}$.

The world map vectors were gathered from the Natural Earth free vector database as linked in Works Cited.

Many of these outlier data points fall on or close to land (landfall within affected area). The average magnitude of these outliers with a neighborhood size of five

was 5.8416 with a median of 5.7. With a neighborhood size of three, they were 5.7786 and 5.7 respectively.

Future Work

One main limitation we faced within the data set, as stated above, was the absence of data for events below a 5.5 on the Richter scale. Adding smaller events would give finer resolution around faults, reveal small fault splays or secondary structures, and decrease the number of false outliers. Additionally, many smaller seismic events may not be tied to large fault activity, and incorporating data such as plate thickness or crustal stress fields could indicate future areas for increased activity that are not yet producing large, easily-measurable earthquakes. Integrating these geophysical layers with the spatial clustering results would help distinguish between genuine anomalous seismicity and outliers caused by sparse sensor coverage or incomplete fault mapping.

Incorporating depth and magnitude into a multi-dimensional clustering framework would further improve the geological relevance of the clusters by differentiating shallow crustal events from deeper subduction-zone earthquakes.

Another meaningful extension would be the addition of temporal analysis. Because earthquakes often occur in sequences rather than as isolated events, examining how clusters evolve over time could highlight periods of increased stress accumulation or activation along specific fault segments. Temporal clustering may also help differentiate isolated intra-plate events from emerging swarms or precursory activity.

Finally, a systematic investigation of the outliers themselves represents a valuable direction for future study. Outliers may correspond to intra-plate earthquakes, induced seismicity (geothermal generation, fracking, reservoirs, mining), unmapped fault structures, or regions with insufficient sensor coverage. Classifying these outliers according to their geological context could provide insight into global seismic hazards outside of well known plate boundaries. However, much of induced seismicity data on location or effect is not public record, and it would be hard to produce a comprehensive data set that verifies results on a large scale, especially for countries where such record-keeping and release of information is not even as "commonplace" as the US.

This deeper analysis of outlier behavior, combined with a more comprehensive dataset would offer a better understanding of global seismic activity.

Works Cited

1. Levandowski, W., R.B. Herrmann, R. Briggs, O. Boyd, and R. Gold. (2018) An updated stress map of the continental United States reveals heterogeneous intraplate stress, Nature Geoscience, 11, pages 433–437.

2. Natural Earth. 1:110m Cultural Vectors, Admin 0 - Countries, naturalearthdata.com/downloads/110m-cultural-vectors/.

3. Styron, Richard, and Marco Pagani. "The GEM Global Active Faults Database." Earthquake Spectra, vol. 36, no. 1_suppl, Oct. 2020, pp. 160–180, doi:10.1177/8755293020944182.

4. US Geological Survey and Abigail Larion. (2016). Significant Earthquakes, 1965-2016, Version 1.