An Exploratory Data Analysis of the CSZ and NSAF from 2011-2021

Swetha Natarajan, N.C State University

Abstract

The Cascadia Subduction Zone (CSZ) is located from Northern Vancouver Island to Cape Mendocino, California. This region of intense tectonic activity is predicted to cause a magnitude 9.0 or greater earthquake that will wipe out most of the north-western seaboard of the United States. These earthquakes occur on inconsistent intervals that range from 200 to 1,000 years. Hence, scientists are unable to make confident predictions about tectonic activity in Cascadia.

The Northern San Andreas Fault is a region of frequent seismic activity and stretches from the Mendocino Triple Junction through the San Francisco Bay Area. Research suggestions that quakes in the CSZ could have triggered earthquakes in NSAF in the past one thousand years.

In order to progress research in this area, the following investigation of CSZ and NSAF provides an understanding of our approach to this eventual earthquake and enhanced predictions on its location, depth, and magnitude. This exploratory analysis of earthquake data from 2011-2021 in both regions will aid scientists in understanding patterns of seismic behavior in both regions and determine whether a causal relationship exists between the two faults.

Background

Subduction zones are areas where two tectonic plates meet, allowing one plate to slide under the other and into the Earth's mantle. The Juan de Fuca plate dives under the North American Plate to create CSZ, as seen in Image 2.

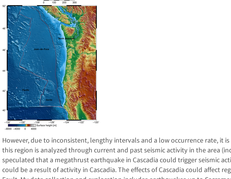
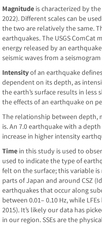


Image 2: Tectonic Plates of the CSZ. Graphic by the National Parks Service

The slipping motion between both of these plates is the cause of "megathrust earthquakes", which have a magnitude of 8.5 or higher. Image 4 shows an aerial view of the side by side plates. On average, CSZ produces an earthquake every two hundred and forty-three years, the most recent being a magnitude 9.0 in January 1700. The area is overdue for an earthquake; we've coined this future earthquake "The Big One" to commemorate its size. If accurate predictions can be made using data from nearby faults and the past, then governments and residents can prepare accordingly.



However, due to inconsistent, lengthy intervals and a low occurrence rate, it is difficult to predict when the next "big one" will occur. Plate movement in this region is analyzed through current and past seismic activity in the area (increases in magnitude, occurrences, and changes in earthquake depth). It is speculated that a megathrust earthquake in Cascadia could trigger seismic activity in the San Andreas Fault, meaning movement in the San Andreas fault could be a result of activity in Cascadia. The effects of Cascadia could affect regions up to Sacramento, California, which lies on the Northern San Andreas Fault. My data collection and exploration includes earthquakes up to Sacramento for the possibility of future analysis of the relationship between the two faults.

Introduction

The characteristics of depth, magnitude, and time can be used to draw conclusions about seismic activity. **Depth** is a measurement of the distance between the earth's surface and the location at which the earthquake begins to rupture (U.S Geological Survey, 2022). This point where the earthquake ruptures is its epicenter. Earthquakes can occur either in the crust or upper mantle of the earth, approximately 800 kilometers into the Earth's surface.

Magnitude is characterized by the maximum motion recorded by a seismograph, and is regarded as the size of an earthquake (U.S Geological Survey, 2022). Different scales can be used to measure magnitude; the Richter scale or the Mercalli scale. Regardless of the scale used, measurements between the two are relatively the same. The Mercalli scale is more reliable for measuring larger earthquakes and the Richter scale is more reliable for smaller earthquakes. The USGS ComCat measurements of magnitude follow the Richter scale. This scale uses a logarithmic interval to determine the amount of energy released by an earthquake, and magnitudes range from approximately 1.0 to 9.0, with 9.0 being the highest. Since it follows a logarithmic scale, seismic waves from a seismogram for a 6.0 earthquake are ten times less than that of a 7.0 earthquake.

Intensity of an earthquake defines the severity of the shaking produced by an earthquake. The intensity of an earthquake, regardless of magnitude, is dependent on its depth, as intensity decreases the greater the distance from the earthquake's epicenter. An earthquake with an increased distance from the earth's surface results in less shaking compared to an earthquake with a lower depth (U.S Geological Survey, 2022). Intensity is also often based on the effects of an earthquake on persons, structures, and the environment..

The relationship between depth, magnitude, and intensity is as follows; the shallower the earthquake and the larger the magnitude, the more intense it is. An 7.0 earthquake with a depth of 70 km will likely result in less damage compared to the same earthquake at a depth of 30 km under the surface. An increase in higher intensity earthquakes may indicate danger for CSZ.

Time in this study is used to observe changes in seismic activity in and around CSZ and NSAF from 2011-2021. **Duration**, a measurement of time, can be used to indicate the type of earthquake observed. Duration can either be the amount of time it takes a fault to crack, or the amount of time shaking is felt on the surface; this variable is not included in our dataset. **Very Low-Frequency Earthquakes** (VLEs) are a new variety of earthquakes discovered in parts of Japan and around CSZ (Ide et al., 2007). The subduction zone has seen LFEs, or low-frequency earthquakes in its southern margin. LFEs are small earthquakes that occur along subduction zones with continuous tectonic tremors for longer periods of time. VLEs are a subset of LFEs with a frequency between 0.01– 0.10 Hz, while LFEs have a frequency > 1 Hz. In Cascadia, LFEs have occurred for 2-3 week long periods every 10-20 months (Plourde et al., 2015). It's likely our data has picked up on a slow slip event (SSE) over the past ten years, but without information on duration, we cannot single out LFEs in our region. SSEs are the physical process in a subduction zone that produces "slow earthquakes" which characterize LFEs and VLEs.

All variables can be used to gauge changes in Cascadia's seismic behavior. An approach to "The Big One" may be indicated with increases in magnitude and occurrence over time, and the types of earthquakes (shallow or slow). Investigating data about the past may point out patterns about future events in this region.

Methods:

Seismic data for this paper was taken from the U.S Geological Survey Earthquake Catalog (USGS). The USGS data was selected because USGS research is used by all levels of government and the private sector; any scholarly publications that go through USGS are required to be public access and are limited to research conducted by USGS or funded by USGS. This ensures precision and validity of data and easy accessibility.

RStudio was utilized in data storage, filtration, and in the creation of data visualizations. RStudio is free, open-source software, useful for exploratory data analysis (EDA). The object-oriented programming language R is efficient to organize and manipulate multiple datasets in a single environment. I considered SAS when determining what software to run my EDA in, depending on the size of my datasets; SAS is efficient for big data while R is not. R can run many packages that allow the user to apply functions on data dependent on the package. The packages "ggplot2", "gridExtra", "dplyr", "shiny" and "leaflet" were used for this report.

ArcGIS Online was used to create additional web maps. ArcGIS is a cloud-based software created by Esri that builds interactive, accessible, online web maps. It offers intuitive analysis tools and specializes in location data and can host multiple types of datasets, from GeoJSON files to .CSV. It is not an open-source software and does not offer a variety of options for visualization output; this is limited to a PDF or a web map.

Datasets were obtained from the USGS Earthquake catalog. This allows the user to filter data by factors (i.e., location, interval of time, minimum or maximum magnitude of earthquake, type of seismic event), and offers KML, .CSV, GeoJSON, QuakeML, and Map & List formats. .CSV files were chosen because of the use of RStudio that can easily read in .csv files with the "read.csv" function. CSZ data was taken with coordinates between 38.000 to 47.000° N and -122.000 to -125.000° W, with magnitudes greater than 2.5. NSAF data was taken with coordinate **Insert coordinates here** Since VLFs have a magnitude around 3, this does accommodate any slow earthquakes. Our data doesn't contain information about the duration of earthquakes relative to magnitude, so we are assuming all earthquakes in our data are shallow earthquakes. Each data set contains data corresponding to 22 different variables and filtered to 14 variables. Variables kept include "time", "latitude", "longitude", "depth", "mag", "magType", "nst", "gap", "dmin", "rms", "net", "place", "type", "magError". Not all kept variables were used in EDA. Many variables removed contained data not pertaining to the purpose of the data, but the validity of it, ie, "magError" and "status". MagError is the standard error for the magnitude, and status indicates whether a datapoint has been reviewed by a human. The core variables of this set include "depth", "mag", "latitude", "longitude", and "time". Additional variables kept were for the purpose of future analyses. However, variables like "nst", "gap", "dmin", "rms", and "net" also hold information regarding distance from the event to the nearest seismic station, the root mean square error, and more on the validity of the data point.

Results

Research generally relies on **hypothesis testing** to draw conclusions regarding a population parameter, such as the mean or median. Most testing does require assumptions of normality. Earthquake data generally follows a **power law distribution**.

Image 4: Power Law Distribution. Graphic by VISIBLE

Power law distributions show relative change in one quantity resulting in a proportional relative power change in another quantity. If a square's length and area followed a power law relationship, then doubling the length would increase the area by a factor of four. Earthquake magnitude and frequency follow a power law distribution, where there is a proportional relationship between magnitude and frequency. As the magnitude of earthquakes in a defined area increases, the frequency of occurrence decreases. In seismology, the Gutenberg-Richter law expresses this relationship as:

$$N = 10^{(a-bM)}$$

$$N = 10^{(a-bM)}$$

where M is magnitude. N is the number of events with a magnitude greater than or equal to M, and a and b are constants.

Figure 1 below displays the distribution of magnitude for each year for CSZ and Figure 2 for NSAF:

Figure 1: Distribution of CSZ Earthquakes from 2011 to 2021

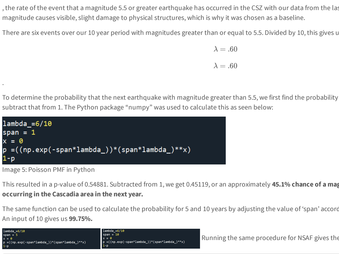
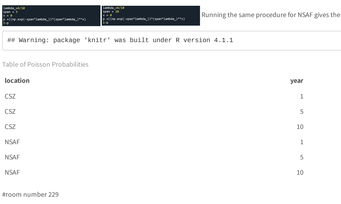


Figure 2: Distribution of NSAF Earthquakes from 2011 to 2021



Both figures show consistent results that the magnitude of each year for both locations follow power law distributions. Hence, tests with normality assumptions will not be applicable here.

To compare magnitude between the two groups over a 10-year period, we can perform a **Wilcoxon Rank Sum test**, also known as the **Mann Whitney U test**, is a non-parametric test used when data are robust and assumes the data is distribution free, based on rank values, and independence between groups.

To compare and predict the number of future earthquake events, we can assume a **Poisson Distribution** and use its **probability mass function (pmf)**. This is function used to determine the probability that a discrete random variable equals some specified value; it can make predictions on the likelihood of an event of a specific magnitude occurring. Earthquakes do occur randomly but over a long period of times they approach a constant rate. The distribution assumes countable, independent events that do not occur simultaneously, and that the event maintains a homogeneous rate of occurrence (an average rate of which events occur can be found).

The Poisson distribution is generally used in earthquake research but presents many issues. Assuming the average rate of earthquakes is consistent over time does not account for external factors that can trigger earthquakes (i.e., eruptions, fracking).

The 'Quiet Period' of 2020 represents this flaw; research has shown COVID-19 lockdown measures, which lead to decreased economic, industrial, and travel activity, accounted for a 50% reduction in seismic noise around the world. Our CSZ data shows a record low 44 earthquakes picked up during this time (McGill University, 2020).

We chose to run the Poisson distribution to make predictions on 1, 5, and 10 year intervals keeping this in mind with both the CSZ, NSAF, and combined data sets.

To begin this calculation we first must find

$$\lambda$$

$$\lambda$$

, the rate of the event that a magnitude 5.5 or greater earthquake has occurred in the CSZ with our data from the last 10 years. An earthquake of this magnitude causes visible, slight damage to physical structures, which is why it was chosen as a baseline.

There are six events over our 10 year period with magnitudes greater than or equal to 5.5. Divided by 10, this gives us

$$\lambda = .60$$

$$\lambda = .60$$

.

To determine the probability that the next earthquake with magnitude greater than 5.5, we first find the probability that it doesn't occur in one year, and subtract that from 1. The Python package *numpy* was used to calculate this as seen below:

```
lambda_=6/10
span = 1
x = 0
p =((np.exp(-span*lambda_))*(span*lambda_)**x)
1-p
```

Image 5: Poisson PMF in Python

This resulted in a p-value of 0.54881. Subtracted from 1, we get 0.45119, or an approximately **45.1% chance of a magnitude 5.5 or greater earthquake occurring in the Cascadia area in the next year.**

The same function can be used to calculate the probability for 5 and 10 years by adjusting the value of 'span' accordingly. An input of 5 gives us **95.02%.** An input of 10 gives us **99.75%.**

```
lambda_=6/10
span = 5
x = 0
p =((np.exp(-span*lambda_))*(span*lambda_)**x)
1-p
```

```
lambda_=6/10
span = 10
x = 0
p =((np.exp(-span*lambda_))*(span*lambda_)**x)
1-p
```

Running the same procedure for NSAF gives the following results:

```
## Warning: package 'knitr' was built under R version 4.1.1
```

Table of Poisson Probabilities

location	year	probs
CSZ	1	45.10
CSZ	5	95.02
CSZ	10	99.75
NSAF	1	32.96
NSAF	5	86.46
NSAF	10	98.16