SIGNIFICANT EARTHQUAKE 1900-2023

Andrés Leonardo Manrique Hernández
Wilberson Osorio Rodríguez

26 de Noviembre de 2024

Repositorio

Abstract

Los datos analizados en este proyecto provienen de un conjunto de datos de terremotos significativos entre 1900 y 2023, obtenidos de Kaggle. La base de datos contiene información sobre diversos atributos de los terremotos, como magnitud, profundidad, ubicación (latitud, longitud) y continente. A través de una exploración de estas variables, el proyecto pretende descubrir tendencias y patrones en la distribución de los terremotos en todo el mundo. El análisis se centra en estadísticas descriptivas, exploración visual de variables clave y la aplicación de pruebas de hipótesis para comparar las características de los terremotos en diferentes regiones.

1 Introducción

Los terremotos son un fenómeno natural que ocurre cuando se produce una liberación repentina de energía en la corteza terrestre, hecho que puede causar vibraciones y sacudidas en la superficie de la Tierra. Esta liberación de energía se produce cuando las placas tectónicas, que son grandes bloques de la corteza terrestre, se mueven y chocan entre sí. Los terremotos pueden variar en intensidad, duración y afectaciones dependiendo del lugar donde se produzcan. [1]
Es por esto que encontramos interesante revisar una base de datos referente a este tema, es decir, una base de datos de terremotos. Y sobre esta base de datos entender más este fenómeno natural.
El presente trabajo busca realizar un análisis estadístico de en el conjunto de datos sísimicos seleccionado.

2 Marco metodológico

Nuestra base de datos proviene de Kaggle, una plataforma de ciencia de datos en línea, y fue recopilada a partir del conjunto de datos **Significant Earthquake Dataset (1900-2023)**, disponible en la plataforma. Este conjunto de datos recoge información detallada sobre los terremotos más significativos ocurridos a nivel global entre los años 1900 y 2023, ofreciendo un panorama amplio sobre los eventos sísmicos más importantes durante este periodo.

Los datos incluyen información crucial sobre los terremotos, como la magnitud, la ubicación geográfica (latitud y longitud), la profundidad del epicentro, la fecha del evento y la región afectada. [3] La base de datos en bruto cuenta con 23 columnas y 37331 registros. No obstante, al depurar, normalizar, estandarizar, y filtrarla de los valores perdidos (*missing data*) obtuvimos una base de datos con 11 variables, **5 cualitativas** y **6 cuantitativas**. Y al eliminar los datos que tenían varias variables en **NA**, nos quedamos con 7139 registros.

Las variables anfitrionas para este proyecto fueron las siguientes:

1 **ID** *(str)* Id del terremoto.

2 **Year** *(int)* Año en que ocurrió el terremoto.

3 **Region** *(str)* Región geográfica donde tuvo lugar el terremoto.

4 **Continent** *(str)* Continente asociado a la región afectada.

5 **Longitude** *(float)* Coordenada geográfica Longitud del epicentro. De -180° a 180°.

6 **Latitude** *(float)* Coordenada geográfica Latitud del epicentro. De -90° a 90°.

7 **Depth** *(float)* Profundidad del epicentro en kilómetros.

8 **Mag** *(float)* Magnitud del terremoto, medida de acuerdo a la escala *MagType*.

9 **MagType** *(str)* Algoritmo usado para calcular la magnitud registrada el terremoto [4].

10 **MagSource** *(str)* Fuente que reportó los datos sobre la magnitud del terremoto.

11 **NST** *(int)* Número de estaciones sísmicas que registraron el fenómeno.

Para obtener detalles más profundos de las variables se recomienda revisar el diccionario de variables ofrecido por la USGS para la base de datos, disponible en [6].
Los terremotos incluidos en este análisis se consideran "significativos" según los criterios del Servicio Geológico de los Estados Unidos (USGS) [6], lo que implica que estos eventos tuvieron un impacto considerable en términos de destrucción, pérdida de vidas o cambios geológicos notables. El análisis a estos datos nos permite explorar la relación entre diferentes variables como la profundidad y la magnitud, así como observar cómo ciertos patrones sísmicos varían entre continentes y regiones del mundo.

3 Análisis descriptivo

3.1 Visualización e interpretación de los datos

La idea es obtener gráficos como histogramas, diagramas de barras, y demás e interpretarlos, para así tener una breve contextualización de la base de datos.

Así pues, para empezar, veamos cómo se ven nuestras variables cuantitativas y para ello, renderizamos a cada variable su respectivo histograma. Adicionalmente, dado que contamos con la suerte de tener las variables *Latitude* y *Longitude*, podemos acceder a una librería de `R` llamada `Maps` y poder ver las distribuciones de los terremotos de nuestra base de datos en el mapamundi.

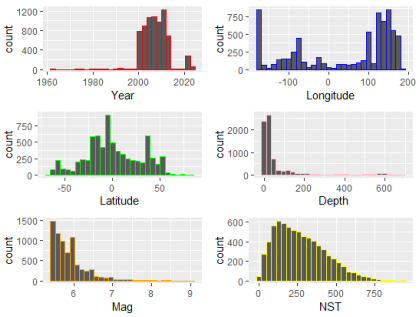

Figure 1: Histogramas de las 6 variables cuantitativas.

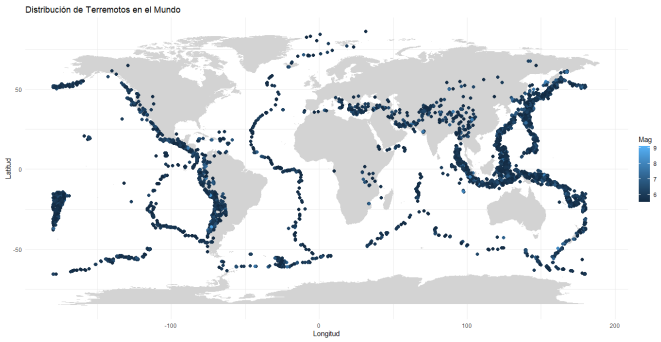

Figure 2: Distribución de los terremotos en el mundo (latitude y longitud).

En la visualización Fig. 1, vemos las distribuciones de las 6 variables cuantitativas que tenemos. Éstas son Año, Latitud, Longitud, NST, profundidad y magnitud del terremoto. Se puede observar que la mayoría de los terremotos con magnitud registrada están entre 5 y 7, mientras que los terremotos con profundidades superiores a los 200 km son raros. Además, la distribución de la latitud muestra más actividad sísmica cerca del ecuador. Este es el punto de 0° en las escalas de la variable *Latitude*. Por su parte, la mayor parte de los terremotos en cuestiones de *Longitude* están en el hemisferio Norte. Así

mismo, la variable *NST*, que representa el número de estaciones sísmicas que reportaron el terremoto, parece tener su media en las 250 estaciones. Es decir, en promedio cuando ocurre un sismo lo detectan 250 estaciones. Por último, los años más activos de los terremotos no es ni tan lejos de nuestro año actual, entre los años 2000 y 2010.

En la visualización Fig. 2, el mapa mundial muestra la distribución geográfica de los terremotos, con las áreas más activas a lo largo de los límites de las placas tectónicas, como el Cinturón de Fuego del Pacífico. Las zonas de mayor actividad incluyen las costas del Pacífico en América, Asia oriental, y el sudeste asiático, reflejando la mayor concentración de terremotos en estas regiones. Las diferencias en las coloraciones de los puntos reflejan la magnitud de los terremotos, a más claro el punto, el terremoto tuvo mayor magnitud.

Revisemos algunos gráficos de cajas y bigotes, con el fin de entender las distribuciones de las variables Magnitud (*Mag*), número de estaciones que reportaron el terremoto (*NST*) y año (*Year*).

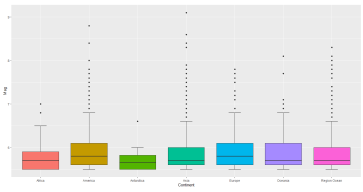

Figure 3: Boxplot de la magnitud categorizado por los continentes.

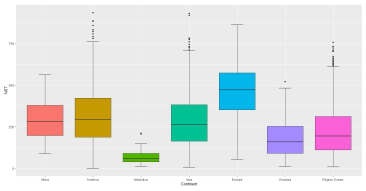

Figure 4: Boxplot de NST categorizado por los continentes.

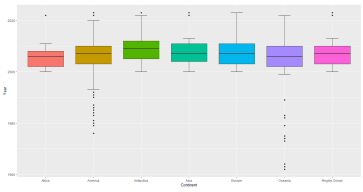

Figure 5: Boxplot de los años categorizado por los continentes.

Este gráfico de caja Fig. 3 muestra la distribución de la magnitud de los terremotos por continente. Hay que tener en cuenta que se hizo un proceso de filtración en la base de datos dado que en un principio no existía la categoría *Region Ocean*, si no que eran diferentes terremotos en lugares oceánicos como el Océano pacífico, indio y atlántico. Entonces, para no tener en cuenta muchas regiones oceánicas, usamos esta categorización. Note que, aunque la mediana de la magnitud es similar entre continentes, América y Asia tienen una mayor cantidad de terremotos de mayor magnitud (más de 8). Antártida no contempla terremotos de magnitudes superiores a 6, a excepción de datos atípicos. África también muestra una menor variabilidad, con la mayoría de los terremotos alrededor de una magnitud de 6.

En este gráfico de caja Fig. 4, se observa la distribución de la variable NST (número de estaciones que registran un terremoto) para cada continente. Podemos ver como Europa destaca con la mayor variabilidad y la mediana más alta, dado que su caja es la más alta, lo que indica que en este continente los terremotos tienden a ser registrados por un mayor número de estaciones, lo que indica a su vez que el terremoto cubre más terreno. En contraste, Antártida muestra la menor variabilidad y un número más reducido de estaciones que registran terremotos. Esto podría deberse a su geografía y menor cantidad de instalaciones de monitoreo probablemente, de igual forma vimos en el mapa Fig. 2 que casi no se ven terremotos por esta zona. América y Asia cuentan con los registros de los terremotos que mayor número de estaciones sísmicas, es decir, terremotos que cubrieron más zonas.

Este último boxplot Fig. 5 muestra la distribución temporal de los terremotos por continente. Se puede ver que los registros suelen tener medias cerca de los años 2007. Por su parte, en Oceanía y América los terremotos se extienden desde aproximadamente 1960, que son algunos de los terremotos antiguos registrados. Antártida y Asia tienen registros más recientes sin muchos eventos anteriores a 1980.

También revisemos otros gráficos como lo son los diagramas de barras y las gráficas de torta o circulares.

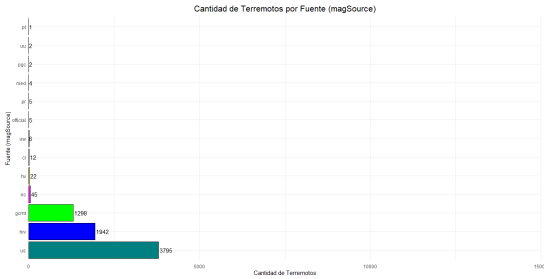

Figure 6: Barra acostada de las fuentes de las magnitudes.

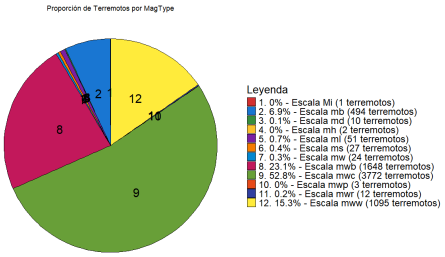

Figure 7: Diagrama circular de los tipos de magnitudes.

El gráfico de barras acostadas Fig. 6 muestra la cantidad de terremotos registrados por distintas fuentes de magnitud. La fuente US (*Station of United State*) es la que ha registrado la mayor cantidad de terremotos, con 3795 registros, seguida de HRV (*Station IU HRV*)es la segunda estación sísmica que más registros tiene, con 1942 terremotos registrados, también resalta GCMT (**Global Centroid Moment Tensor**) con 1298 registros. Las demás fuentes contribuyen en menor medida, con algunas registrando menos de 50 eventos, pero que está disponible para observar en el gráfico con sus respectivo números de registros, si es de interés alguno en particula.

El gráfico de torta Fig. 7, nos da a entender que la escala más utilizada para registrar terremotos es claramente mwc, representando el 52.8% de los terremotos registrados. Otras escalas comunes incluyen mwb (23.1%) y mw (0.3%). También se puede ver por ahí resaltando la escala mb (6.9%). Esto indica que las cuaatro escalas principales dominan el registro de terremotos, mientras que las demás se usan en un porcentaje mucho menor.

4 Análisis De Componentes Principales (PCA)

El siguiente es un análisis de componentes principales **(PCA)** [2] que busca identificar cómo se comportan los terremotos registrados y, de alguna forma, categorizar los grupos que se crean entre las variables, mostrándonos la familiaridad que tienen entre sí y cómo esto nos ayuda a saber las características de ciertos terremotos.

4.1 Métricas de calidad del modelo PCA

Para obtener un buen análisis de componentes principales, es necesario aprobar dos pruebas sobre los supuestos de los datos;

- **KMO**: La primera medida de calidad del PCA es que el criterio de Kaiser-Meyer-Olkin (KMO) debe ser superior a 0.7. Esto demuestra que existen altas relaciones entre las variables y el análisis sería más idoneo.
- **Esfericidada de Bartlett**: La segunda prueba es la prueba de esfericidad de Bartlett, esta prueba prueba la correación de las variables entre sí y su idoneidad para justificar el uso del método PCA. La prueba tiene las siguientes hipótesis:

$$H_0 : \Sigma = I \qquad vs \qquad H_1 : \Sigma \neq I$$

  Esto básicamente dice que, de aprobarse la hipótesis nula (H_0) las variables no están correlacionadas, y por tanto, no sería ideal hacer un análisi de componentes principales. Por lo tanto, lo que buscamos es rechazar H_0 en favor de H_1, ya que así, aseguramos que existe una correlación importante entre las variables apta para el PCA.

Sobre nuestro modelo
En nuestro Análisis obtuvimos las siguientes métricas:

$$KMO = 0.42$$

$$P_{value} \approx 0 \qquad \text{(P-valor de la prueba de Bartlett)}$$

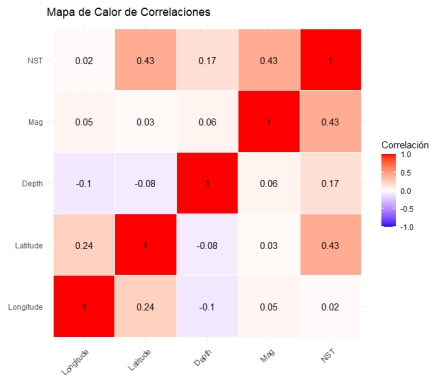

Figure 8: Matriz de correlaciones.

Es decir, por un lado, la prueba de esfericidad de Bartlett se rechaza con una significancia del 5% dado que $P_{value} < 0.05$ y así, en efecto las variables están lo suficientemente relacionadas para hacer un PCA, esto también lo podemos ver en nuestro matriz de correlaciones en la Fig. 8. No obstante, por el otro lado de la moneda, el KMO fue menor que 0.7. Tuvimos problemas para mejorarlo. A lo máximo que se logró llegar fue a un $KMO = 0.58$. Ahora bien, esto gracias a varias transformaciones de las variables iniciales, lo que indica que obtuvimos un modelo que no era sencillo de interpretar. Por lo tanto, teniendo en cuenta la filosofía de parsimonia, decidimos quedarnos con el modelo sin transformaciones, aun conscientes de que el KMO no era el correcto.

4.2 Varianza explicada

Teniendo en cuenta que, aunque nuestros datos no son lo suficientemente idóneos para realizar este análisis de componentes principales, decidimos realizar este análisis exploratorio con el fin de interpretar lo que nuestro modelo puede decir de los datos. Por lo tanto, inicialmente, revisamos la cantidad de varianza explicada por componentes:

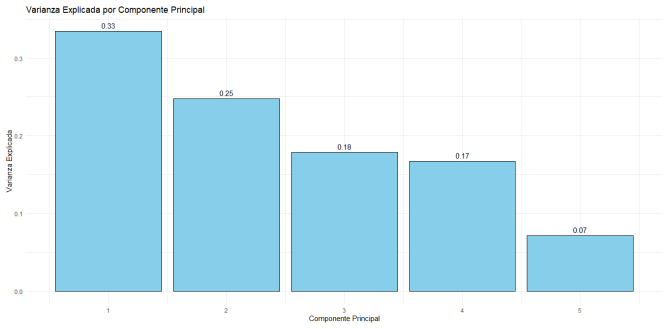

Figure 9: Varianza explicada por los componentes del PCA.

Podemos notar que los dos primeros componentes explican cerca del 57% de la varianza total de los datos. Esta particularmente no es una gran explicación de la variabilidad de estos. Sin embargo, si consideráramos el tercer componente la varianza explicada sería al rededor del 70%.

Para el propósito de este documento, se realizará el análisis en dos dimensiones, por lo tanto, nos conviene sólo usar **las dos primeras componentes**. Sólo como dato extra, en el repositorio de GitHub de este

informe se encuentra a detalle un Notebook y un código de R en el que sí se hizo la apreciación de esa tercera componente a fin de conocer bien las agrupaciones de las variables.

4.3 Componentes PCA1 y PCA2

Al analizar en un plano cartesiano de dos dimensiones nuestras componentes 1 y 2, obtenemos la siguiente salida.

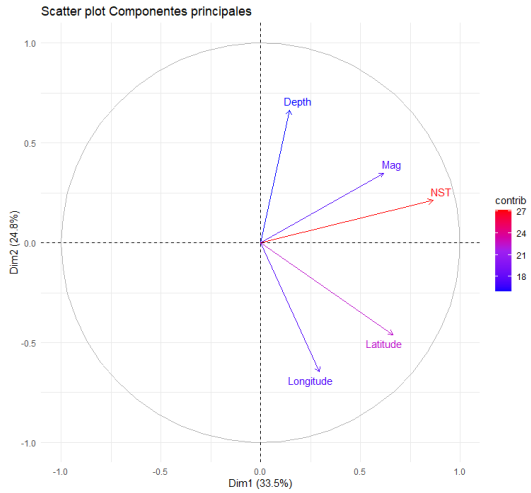

Figure 10: Variables en los componentes 1 y 2.

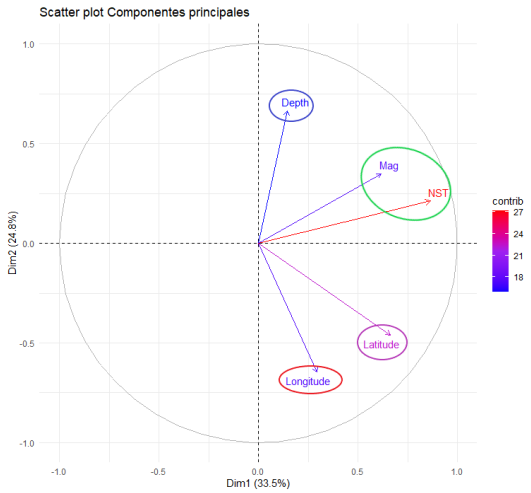

Figure 11: Reducción dimensional.

El gráfico de la Fig. 12 sugiere que las variables *Mag* y *NST* están mejor representadas por la primera componente. Mientras que las variables *Depth* y *Longitude* lo están por la segunda componente. La variable *Latitude* no se explica bien desde ninguna de estas componentes.
Por otra parte, la Fig. 13 nos brinda la posibilidad de disminuir la dimensionalidad y pasar de 5 grupos a 4, dado que la magnitud de un terremoto se puede relacionar con el número de estaciones que reportan este fenómeno. Tiene sentido toda vez que a terremotos con mayor intensidad llegan a más lugares las señales de alerta y así más estaciones sísmicas lo reporta.

4.4 Dispersiones de los terremotos por continentes

Es interesante saber cómo se distribuyen los terremotos en este gráfico de PCA, lo que queremos es asociar a los terremotos de cada continente y vincular estos terremotos a las variables de nuestro análisis de componentes principales.

Figure 12: Distribución de los terremotos por continentes en el plano PCA

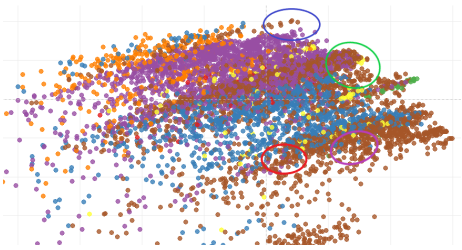

Figure 13: Grupos de variables asociados a los terremotos por continentes.

De estos gráficos podemos observar que los terremotos que tienen mayor grado de magnitud y también mayor número de estaciones que lo reporten son los terremotos que están en el primer cuadrante del plano del PCA. Estos son terremotos del grupo oceánico (recordemos que este grupo se le asigno a los terremotos que ocurrían en los océanos y no precisamente sobre la superficie de un país o región). Según National Geographic [5] los terremotos con las mayores magnitudes reportados ocurrieron una región de costas y océanos denominada *Cinturón de fuego* que como podrá ver en la Fig. 14 se parece mucho a nuestra distribución de los terremotos en el globo terráqueo de la Fig. 2

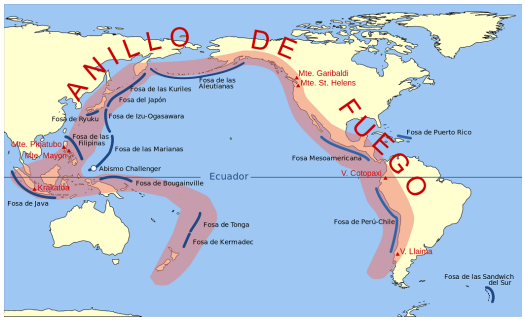

Figure 14: Cinturon de Fuego.

Por otro lado, podemos evidenciar que los terremotos del continente asiático tienen la particularidad de que su epicentro se encuentra a mayor profundidad.

5 Conclusiones

Todo el recorrido de este informe nos ha dejado análisis interesantes sobre los terremotos significativos ocurridos entre los años 1900 y 2023, tales como el comportamiento que tuvieron ciertos terremotos, sus magnitudes, sus distribuciones en el globo terráqueo, etc. Para realizar el análisis se tuvo que limpiar y normalizar los datos. Al final, después de toda la filtración se logró depurar y obtener 7139 registros limpios, sin NA's. También se decidió trabajar con variables que incluían información clave como la magnitud, profundidad y ubicación geográfica de los terremotos. Todo esto nos permitió identificar tendencias, como la mayor actividad sísmica cerca del ecuador y a lo largo de los límites de placas tectónicas, destacando una región en particular denominada el Cinturón de Fuego del Pacífico.

El análisis de componentes principales (PCA) fue una herramienta útil para reducir la complejidad de los datos y descubrir relaciones importantes entre las variables. Aunque enfrentamos limitaciones en algunas métricas o medidas de calidad, como el índice KMO, logramos identificar cómo la magnitud de un terremoto está estrechamente relacionada con el número de estaciones sísmicas que lo detectan. También notamos que estas características varían notablemente entre continentes y zonas oceánicas.

En conclusión, este trabajo contribuyó a una comprensión más detallada de los fenómenos sísmicos, así como también destacó el valor de los análisis estadísticos y visuales para interpretar la base de datos, que en principio era mucho más compleja. Estos hallazgos pueden ser útiles para fortalecer las estrategias de gestión del riesgo ante terremotos, considerando que es un hecho que las zonas que hacen parte del cinturon de fuego están propensas a experimentar terremotos de mayor magnitud. Así como también nos podemos llevar una estimación de la magnitud y fuerza de los terremotos según el número de estaciones sísmicas que reporten el terremoto.

References

[1] Ecoexploratorio: Museo de Ciencias de Puerto Rico. ¿qué son los terremotos?, 2024. Accessed: September 21, 2024.

[2] IBM. Principal component analysis (pca), 2024. Accedido: noviembre 26, 2024.

[3] Jahaidul Islam. Significant earthquake dataset (1900-2023), 2023. Accessed: September 2024.

[4] José Antonio Peláez. Sobre las escalas de magnitud. *Enseñanza de las Ciencias de la Tierra*, 19(3):267–275, 2011.

[5] Redacción National Geographic. Las 3 regiones con más terremotos del mundo. *National Geographic en Español*, 2023.

[6] United States Geological Survey. Anss comprehensive earthquake catalog (comcat). 2020.