BEDU

DATA SCIENCE

REPORTE DE PROYECTO

Bloque 3: Procesamiento de Datos con Python

(Equipo 2)

***Gómez Barraza**, Karla Daniela*
***Guevara Moreno**, Fernando*
***Muñoz Esparza**, José Cruz*
***Rodríguez Rivera**, Gil Estéfano*
***Rosado Martínez**, Ana Elizabeth*

13 de agosto de 2021

Índice

Resumen

En este reporte se presenta una descripción de los componentes principales del proyecto de este bloque del curso. El proyecto consiste en la exploración del tema de los terremotos en el mundo con la información del dataset ***Significant Earthquakes, 1965-2016*** (obtenido de https://www.kaggle.com/usgs/earthquake-database). Esta exploración da respuesta a preguntas generales como cuál es la cantidad de terremotos en distintas regiones del mundo, la magnitud de éstos, así como cuál es la cantidad de sismos anuales. También se abordan preguntas de profundización, como la cantidad de sismos que habrá en los próximos años o la cantidad de sismos en México en el año 2022. Se presenta también una breve reflexión sobre la relación entre la distribución geográfica de los terremotos y las placas tectónicas.

El **Notebook** se encuentra en **Google Colab**: https://colab.research.google.com/drive/13T6rYJlOyld0x1NB_0-TiGYD4QEANFJn?usp=sharing.

Los **códigos** se encuentran en el siguiente repositorio de **GitHub**: https://github.com/gilesitorr/DataScience3_Bloque3.

1. Introducción

El tema de los **terremotos** se puede estudiar desde diversas perspectivas. Es posible determinar las zonas del planeta con más sismos así como predecir la cantidad de sismos anuales en cada región. También se puede predecir la cantidad de sismos que habrá en próximos años. Esto puede ser de utilidad para elaborar planes de evacuación especializados en las zonas con mayor frecuencia de sismos así como de mayor magnitud. Sin embargo, según el Sistema Sismológico Nacional (de México) (s.f.), en la actualidad no hay una manera conocida para predecir cuándo habrá un sismo, mucho menos su magnitud.

Se busca estimar las regiones de riesgo sismológico (en base a la magnitud de los terremotos y su frecuencia). Esto es consecuencia de que no hay métodos actuales para predecir cuándo ocurrirá un sismo, pero las instituciones de protección civil no pueden verse limitadas por eso, entonces la manera de realizar planes de prevención es identificar patrones estadísticos en base a las observaciones pasadas de terremotos.

El primer objetivo que se identifica es determinar las zonas con más sismos anuales (que, a priori, parece una buena unidad temporal para el análisis) así como

las zonas con sismos más “fuertes”. El segundo objetivo es predecir la cantidad de sismos extrapolando la información histórica de los sismos.

En la **Sección 2** se enuncian los conceptos que son relevantes para comprender conceptual y contextualmente algunos elementos del proyecto, además, se incluyen las preguntas guía del proyecto. En la **Sección 3** se describen las partes principales en el desarrollo del proyecto (sobretodo en la limpieza del dataset usado). En la **Sección 4** se explica el proceso seguido para responder las preguntas guía del proyecto así como se discuten brevemente los descubrimientos hechos con el dataset. En la **Sección 5** se mencionan las posibles vertientes que puede tener la investigación actual para futuros proyecos.

2. Marco teórico

De acuerdo a Wikipedia (s.f.), un terremoto (o sismo) es un fenómeno donde la corteza terrestre se sacude de manera brusca y pasajera al liberar energía en forma de ondas sísmicas. Estas ondas generalmente ocurren por la actividad de las fallas geológicas así como por el movimiento de las placas tectónicas, al igual que por procesos volcánicos o impactos de asteroides. El punto de origen de un terremoto se conoce como foco o hipocentro, mientras que el punto de la superficie que se encuentra sobre el hipocentro se conoce como epicentro.

Según el Sistema Sismológico Nacional (s.f.), se tienen diversas escalas para medir el tamaño o el impacto de un temblor. La magnitud de un temblor está relacionada con la energía liberada en forma de ondas sísmicas que se propagan a través del interior de la Tierra. Esta energía -que, a su vez, sirve para determinar la magnitud- se calcula a partir de algunas características de las ondas así como la distancia entre el epicentro y la estación de medición (los aparatos usados para esas mediciones se conocen como sismógrafos). La escala de magnitud se obtiene a partir de los registros obtenidos por sismógrafos. Actualmente ya no se usa la escala de Richter original; las magnitudes que se usan actualmente son la magnitud de coda (Mc), la magnitud de energía (Me) y la magnitud del momento sísmico (Mw), entre otras. Cada una de las escalas tiene sus ventajas y limitaciones. Como mención adicional, otra cantidad usada es la intensidad del sismo, que se asigna en función de los daños o efectos provocados al ser humano y sus construcciones.

2.1. Preguntas

En base a los objetivos que se plantearon y a la investigación, se pueden determinar algunas preguntas guía.

Preguntas generales:

1. ¿Cuántos sismos hay anualmente?
2. ¿Cuál es la relación entre magnitud y frecuencia?
3. ¿Cuáles son las regiones donde hay sismos más frecuentemente?
4. ¿Cuáles son las regiones con más sismos “fuertes” (en términos de magnitud)?

Preguntas de profundización del tema:

1. ¿Cuántos sismos habrá en los próximos años?
2. ¿Hay una relación entre la magnitud o la profundidad de los sismos y su región?
3. ¿Cuáles son las regiones con los sismos de menor y mayor profundidad?

Pregunta específica:

1. ¿Cuántos sismos habrá en México en el 2022?

2.2. Soluciones anteriores

Para dar contexto al proyecto actual, se discuten brevemente algunos proyectos que sirven como referencia.

Como se ve en el blog de Aaron Lee (2020), una de las maneras más comunes de graficar los sismos en un mapa es colocando un punto sobre la ubicación del epicentro. De la misma manera, si se desea detallar, se puede jugar con el tamaño y el color de los puntos para indicar su magnitud. Esta perspectiva sugiere fuertemente las zonas donde hay más sismos detectados incluso si se presentan pocos puntos. Sin embargo, cuando se tienen muchos puntos en un mismo mapa, se pierde la noción visual de la frecuencia de los sismos.

Al priorizar la noción de frecuencia, es necesario restar detalle a otros rasgos, como la magnitud del sismo. Esto puede sortearse dependiendo del interés de la investigación. Es decir, puede determinarse el rango de escalas a estudiar y, una vez seleccionados los registros de la base de datos que cumplen con ese criterio, se puede realizar un heatmap para presentar la información. El Government of Canada (s.f.) en su página oficial presenta unas gráficas que pueden servir de ejemplo (aunque presentan información que se relaciona con la frecuencia de los sismos pero también incorpora otras características al estudio).

La intención con el heatmap de las frecuencias en función de la ubicación es que, dada una serie de tiempo de la cantidad de sismos con la que se hagan predicciones, puede servir para predecir la cantidad de sismos en cada región para un año futuro. Esta puede ser una manera de darle la vuelta a la limitación de no poder predecir cuándo ocurrirá un sismo. Para un análisis similar pero con más detalles que lo propuesto, puede consultarse el trabajo de Kagan, Y. (2009).

3. Desarrollo del proyecto

Con el objetivo de abordar la problemática, se buscó un dataset pertinente. De **Kaggle** se obtuvo el dataset titulado ***Significant Earthquakes, 1965-2016***, proporcionado por el Servicio Geológico de los Estados Unidos. El dataset se encuentra en (https://www.kaggle.com/usgs/earthquake-database).

El dataset contiene 21 columnas, entre las cuales se encuentran la fecha, hora, localización y magnitud de sismos reportados a nivel mundial con magnitudes mayores a 5.5 desde 1965 hasta 2016.

3.1. Análisis Exploratorio de los Datos

Primero se lee el dataframe. Se utilizaron funciones métodos como *shape, head, tail, dtypes e info* para tener una noción de qué contiene el dataset. Las preguntas guía a tener en mente y sus respuestas se enlistan a continuación:

1. **¿El conjunto de datos que tengo realmente me sirve para responder algunas de las preguntas que me planteé?** Sí, pues cuenta con una columna de fecha, otras con longitudes y latitudes (para poder tener idea de la localización geográfica) así como una columna de magnitudes.

2. **¿Qué tamaño tiene mi conjunto de datos? ¿Serán datos suficientes?** El dataset completo tiene una dimensión de 23,412 registros y 21 columnas. Considerando que las columnas que son de interés están completas, se puede ver que son datos suficientes.

3. **¿Qué columnas tengo y qué información tengo en cada una de esas columnas?** Las columnas relevantes son *Date* (fecha del evento), *Latitude* (latitud), *Longitude* (longitud), *Type* (si el temblor detectado es un terremoto o fue causado por algún otro tipo de evento), *Magnitude* (magnitud del sismo), *Magnitude Type* (qué sistema para determinar la magnitud se usó) y *Depth* (la profundidad del hipocentro).

4. **Los nombres que tienen mis columnas, ¿son el nombre más apropiado?** Sí, entre los nombres y los datos, es sencillo comprender la base de datos.

5. **¿Qué tipos de datos tengo en cada columna? ¿Parecen ser el tipo correcto de datos? ¿O es un tipo de datos "incorrecto"?** Date tiene formato de string, entonces será oportuno cambiar el formato a fecha. Latitude, Longitud, Magnitude y Depth tienen formato de float, que es apropiado. Tanto Type como Magnitude Type tienen formato de string, que tiene sentido, sin embargo, Type puede usarse para un filtro para sólo analizar terremotos.

6. **Si selecciono algunas filas al azar y las observo, ¿estoy obteniendo los datos que debería? ¿O hay datos que parecen estar "sucios" o "incorrectos"?** Hay muchos NaNs, pero no en las columnas que son interesantes. Las columnas irrelevantes se quitarán para el análisis posterior.

3.2. (*Opcional*) Automatización y APIs

El proceso de lectura de datos puede hacerse usando el API de *Kaggle*. Para eso se requiere descargar el paquete *kaggle* para Python. Luego se obtiene el API key en el perfil de *Kaggle*. Se consultan los datasets y se escoge el dataset de interés. Luego se descomprime y listo, se tiene libre el archivo csv listo para leerse. La información relevante se puede encontrar en el Notebook.

3.3. Limpieza de datos y agregaciones

Una vez que se lee el dataset y se tiene una noción general de lo que contiene, puede procederse con la limpieza de datos.

Primero se explora la cantidad de NaNs de cada columna, para determinar si hay columnas inservibles. Se descartaron 8 columnas (de las cuales, la que menos NaNs tenía, contaba con 6,060 NaNs, de los 23,412 datos totales). Con esta limpieza original, sólo quedó una columna con 3 NaNs, lo que se hizo es que se aplicó el método *drop* sobre las filas que tenían esos registros con NaNs, quedándonos al final con un dataset con 23,409 registros. Posteriormente, se reindexó el dataset.

Se hizo una breve revisión sobre qué tan dispersos estaban los datos. Además, se revisó con la columna *Type*, cuántos registros pertenecían a terremotos y no a otro tipo de fenómeno: se observó que 23,229 registros de la base de datos original son terremotos. Por lo que, incluso luego de la limpieza del dataset, puede pensarse en un futuro filtro para sólo trabajar con terremotos y aún así tener una cantidad importante de datos.

3.4. Transformación, filtración y ordenamiento de datos

Continuando el trabajo de la subsección anterior, se convirtieron las strings de la columna de fechas en un formato de fecha, para esto se hizo uso de la estructura *try-except*, pues no había un único formato en las strings.

Se filtraron los registros que fueran terremotos. Además, se omitieron algunas columnas que no eran relevantes dentro del contexto del estudio. Las columnas que se usarán en adelante son las que contienen fechas, latitud, longitud, magnitud, tipo de magnitud y profundidad del sismo. A partir de este dataset resultante, se

respondieron las preguntas planteadas en la sección 2. En la sección 4 se presentan desarrollor para responder las preguntas y las respuestas y reflexiones obtenidas.

4. Resultados

4.1. Preguntas generales

¿Cuál es la relación entre magnitud y frecuencia?

Antes de responder cualquier otra pregunta, se hizo un histograma para visualizar la distribución de frecuencias con las que se dan los sismos de algunas magnitudes. Se interpoló sobre esta distribución una función con un decaimiento exponencial de la forma $a + be^{-cx}$. Con esta función se pudo estimar la cantidad de sismos de magnitudes menores a las que se encuentran en el dataset (es decir, menores a 5.5). Luego se estimó las proporciones de cada cantidad con respecto al número de sismos con magnitudes entre el rango de 5.5 y 6.0.

Se notó que cada par de valores con diferencia de un punto completo de magnitud, tienen frecuencias que difieren una potencia de 10).

De hecho, graficando el logaritmo (base 10) de las frecuencias de los sismos en función de la magnitud y ajustando una recta al conjunto de frecuencias medidas, se encontró una recta con pendiente $-1{,}15$. Esta recta sugiere que, en efecto, la frecuencia de los sismos de una magnitud M es diez veces mayor que la de los sismos con magnitud $M+1$ y 100 veces mayor que la de sismos con magnitud $M+2$.

Esta relación se conoce como la **Ley de Potencias de Gutenberg-Richter**. Esta relación es independiente de la región.

Incluso el ajuste de la exponencial de antes sigue esa ley de potencias. El ajuste exponencial usado al inicio fue:

$$2{,}29 + (9{,}81 \times 10^9)e^{-2{,}32x} \sim 10^{10} * (e^{2{,}32})^{-x} \approx 10^{10}10^{-x} = 10^{10-x}$$

Note que, luego de hacer algunas aproximaciones a ese ajuste exponencial, se descubre que todo el tiempo se estuvo trabajando con potencias de 10, pues al sacar el logaritmo base 10 de la última expresión, se obtiene una recta que va como $10 - x$, que es de nuevo una recta con pendiente -1 (pues la ordenada no es tan relevante en este caso, porque su valor sólo depende del tamaño de la muestra que se estudie). Que es la Ley de Gutenberg-Richter.

Con esto se concluyó que la estimación de los sismos totales se sigue directamente de estimar los sismos del dataset, por lo que se puede omitir tal paso en el análisis futuro.

¿Cuántos sismos hay anualmente?

Ahora es posible centrarse sólo en hacer una serie de los sismos totales por año (únicamente con los datos que hay en el dataset). Para eso se necesita de un dataframe con las fechas con formato y las magnitudes.

De las fechas, se obtiene el año y se grafican las ocurrencias de sismos con magnitud superior a 5.5 por año.

Puede verse que la cantidad de terremotos con magnitudes mayores a 5.5 ha aumentado en tendencia, pero con datos anuales, parece haber mucho ruido. **OJO**: En el contexto de las eras geológicas de la Tierra (que transcurren tras decenas o centenas de millones de años), esta tendencia no significa que va a haber más terremotos que hace 100 millones de años. Esta tendencia es algo puramente local y debe interpretarse sólo como un aumento relativo a las décadas recientes y no como un incremento global en la actividad sísmica de la Tierra.

Una mejora posible sería hacer series pero con periodos de tiempo más amplios, como cada cinco años, para tener una noción más clara de la tendencia.

¿Cuáles son las regiones donde hay sismos más frecuentemente?

Se realizó un histograma 2d (que es el término apropiado para un *heatmap de frecuencias*). En este histograma se colocaron las longitudes en el eje horizontal y las latitudes en el vertical. Los *bins* del histograma se escogieron de tal modo que fueran recuadros de 10×10 en el espacio de la longitud $\times$ latitud. Esto para tener suficiente detalle sobre las distintas regiones pero sin perder claridad sobre las zonas con mayor frecuencia de terremotos.

Se observó que la región con más terremotos detectados es la de los cuerpos oceánicos que colindan con Japón y Oceanía así como alrededores. Además, hay que notar que toda la costa asiática con el Oceano Pacífico es muy activa, en comparación con la actividad sísmica de los otros continentes.

¿Cuáles son las regiones con más sismos “fuertes” (en términos de magnitud)?

El sismo de México del 2017 tuvo una magnitud de 8.2 en la escala Mw y, el de 1985, una magnitud de 7.1. Si bien, el impacto de un sismo depende de su cercanía a un centro urbano, puede considerarse que un sismo de magnitud superior o igual a 7.0 ya puede ser muy peligroso. Por lo tanto, se considerará como "*fuerte*.ª todo sismo con una magnitud superior o igual a 7.0.

En este pregunta se siguió un proceso muy semejante al de la pregunta anterior, pero filtrando antes a todos los filtros con una magnitud superior a 7.0.

En el histograma 2d se observó que la zona con más terremotos fuertes es la misma que la que tiene más terremotos.

Este resultado soporta y refuerza el estudio que se hizo al inicio de la sección, sobre la proporción de la cantidad de sismos. Se ve que, en efecto, tiene sentido suponer que, a mayor cantidad de terremotos de alguna magnitud, puede esperarse una mayor cantidad de las demás magnitudes.

4.2. Preguntas de profundización

¿Cuántos sismos habrá en los próximos años?

Para responder esta pregunta, se puede recuperar la serie temporal de los años y las magnitudes de los terremotos y ajustarle un modelo estocástico.

Se hizo un ajuste en RStudio para determinar los mejores coeficientes del modelo ARIMA para ajustar a la serie de tiempo. Con un modelo ARIMA(0,1,1) se hicieron predicciones sobre la cantidad de sismos mundiales con magnitud superior a 5.5 desde el 2017 hasta el 2022.

¿Hay una relación entre la magnitud o la profundidad de los sismos y su región?

Antes de analizar si hay correlación, se revisó la distribución geográfica de los terremotos contra su latitud o longitud.

La distribución en latitud es aproximadamente normal, sin embargo, la de la longitud no. Esto puede parecer problemático, pero en realidad hay que recordar que esa parametrización de la esfera terráquea (en términos de latitud y longitud) sólo es una convención, por lo que si *"se pegan"* los extremos de la longitud, se puede ver que sí hay una distribución más o menos normal (pegando los extremos del histograma y cortando en el cero).

Esta manera de recortar y pegar el histograma se consigue sumando 360 a los extremos con latitud menor a 0. Esto es porque la columna que pertenece a -180 debe encontrarse con la columna de 180.

Luego, revisando la matriz de correlación entre la latitud, la longitud, la profundidad y la magnitud, se notó que no hay una correlación, en realidad. Lo mismo ocurrió utilizando la longitud reparametrizada en lugar de la longitud original. Para verificar si hay patrones a pesar de la correlación, se graficó cada variable contra la otra.

Al hacer un *scatter plot* de los terremotos en su ubicación geográfica (es decir, colocando puntos por cada terremoto en el expacio longitud $\times$ latitud), se observó que las regiones con más terremotos parecen formar líneas. ¿Qué significan esas líneas donde hay más terremotos?. Esas líneas resultaron ser sumamente similares

a las placas tectónicas. Con eso en mente, al agregar la profundidad al *scatter plot*, se puede estudiar de qué manera interactúan las placas (se hunden, sólo hay rozamiento, etc.).

Si bien, lo que encontramos no es ciencia nueva, sí es un ejemplo de cómo los datos pueden contar historias inesperadas.

Lo que se encontró se conoce como una falla (como la de San Andrés, que pasa por las Californias). Muchos terremotos son causados por el movimiento de las placas y su interacción entre ellas. En resumen, los terremotos que vemos en el *scatter plot* son causados por el rozamiento o el hundimiento de las placas tectónicas.

La interpretación de los terremotos y las placas tectónicas recontextualiza las gráficas de la latitud o la longitud contra la profundidad o la magnitud. Las gráficas indicaron que en torno a la antípoda del Meridiano de Greenwich (la longitud 180) y en torno al Ecuador (la latitud 0), se construye una distribución aproximadamente normal de las magnitudes y las profundidades de los terremotos en el mundo. De hecho, si se observan con detenimiento, aunque los picos de las gráficas de la profundidad son más definidos, las gráficas que usan la magnitud presentan unos picos en los mismos rangos de longitud o latitud, según sea el caso.

Esa distribución tiene sentido pues en esa zona se encuentra el **Cinturón de Fuego del Pacífico**. Una región que sigue las costas del Oceano Pacífico que cuenta con las zonas de subducción (cuando una placa tectónica se hunde debajo de otra) más importantes: eso da sentido a que ahí haya los terremotos más fuertes y más profundos.

En respuesta a la pregunta inicial, no hay una correlación directa entre la ubicación geográfica y la profundidad o magnitud de los terremotos (como se observó en las matrices de correlación). Sin embargo, se puede ver que sí hay una relación y hasta una explicación para la ubicación de la mayoría de los terremotos.

4.3. Pregunta específica

¿Cuántos sismos habrá en México en el 2022?

Para realizar este primer análisis, se encerró a México con un cuadrado para determinar, aunque sea de forma burda, la cantidad de sismos que ha habido dentro de México desde 1965 hasta 2016. Se filtraron los datos que se encontraran dentro del recuadro de México aproximado y se contaron.

Con esos datos, se contó la proporción de los sismos en México en comparación al resto del mundo. Encontrando que aproximadamente el $2\,\%$ de los terremotos mundiales ocurren en México.

Con esta proporción y las predicciones hechas anteriormente para la cantidad de sismos mundiales entre 2017 y 2022, se estimó que en el 2022 habrán 10 terremotos

con magnitud superior a 5.5 en territorio mexicano (este valor se encuentra delimitado con intervalos de confianza que indican que habrá entre 7 y 13 temblores).

5. Discusión

En futuros estudios, se puede continuar la línea de investigación sobre las placas tectónicas pero ahora con un dataset de volcanes o de tsunamis, por ejemplo. Esto para mapear con mayor fidelidad los límites de las placas.

Por otro lado, con la información de las proporciones de sismos de todas las escalas así como con la estimación de sismos fuertes, se puede hacer un análisis sobre la cantidad de sismos de todas las magnitudes en México para ciertos años. Además, con un *bootstrap* se puede estimar un error estándar para la proporción de sismos de México contra los mundiales. Por otra parte, puede sofisticarse la delimitación del territorio Mexicano para contar la cantidad de terremotos que ocurrieron dentro de éste.

Un proyecto que puede conjuntar todo lo que se menncionó anteriormente sería repetir los análisis anteriores pero en territorios que no son "cercanos" a las fallas de las placas tectónicas. Es decir, cuáles son las probabilidades y las causas de los terremotos que ocurren "fuera" de esas zonas críticas.

Referencias

- SISTEMA SISMOLÓGICO NACIONAL. (s.f). *Preguntas frecuentes*. Recuperado el 28 de julio del 2021 de http://www.ssn.unam.mx/divulgacion/preguntas/.
- WIKIPEDIA. (s.f). *Terremoto*. Recuperado el 28 de julio del 2021 de https://es.wikipedia.org/wiki/Terremoto#:~:text=Un%20terremoto%E2%80%8B%20(del%20latín,terrestre%20producida%20por%20la%20liberación.
- LEE, A. (2020). Plotting USGS Earthquake Data with Folium. Recuperado el 29 de julio del 2021 de https://levelup.gitconnected.com/plotting-usgs-earthquake-data-with-fol
- GOVERNMENT OF CANADA. (s.f.). *Simplified seismic hazard map for Canada, the provinces and territories*. Recuperado el 29 de julio del 2021 de https://seismescanada.rncan.gc.ca/hazard-alea/simphaz-en.php.
- KAGAN, Y. (2009). *Testing long-term earthquake forecasts: Likelihood methods and error diagrams*. Geophysical Journal International. 177. Recuperado el 30 de julio del 2021 de https://watermark.silverchair.com/177-2-532.pdf.

- WIKIPEDIA. (s.f). *Ley de Potencias de Gutenberg-Richter*. Recuperado el 9 de agosto del 2021 de https://es.wikipedia.org/wiki/Ley_de_Gutenberg-Richter.
- WIKIPEDIA. (s.f). *Cinturón de Fuego del Pacífico*. Recuperado el 9 de agosto del 2021 de https://es.wikipedia.org/wiki/Cinturón_de_Fuego_del_Pacífico.