Agrupación espaciotemporal de sismos ocurridos en Colombia

RESUMEN

Colombia, debido a su complejo contexto geológico, experimenta más de 2000 sismos cada mes, los cuales pueden tener un impacto significativo en centros poblados, infraestructuras y la seguridad de sus habitantes. Este alto nivel de actividad sísmica subraya la importancia de analizar los catálogos de sismicidad, donde la agrupación de sismos puede ofrecer valiosas aproximaciones para modelizar la actividad sísmica, contribuyendo así a una gestión de riesgos más efectiva y a la planificación preventiva.

En este trabajo, abordamos el desafío de la agrupación espaciotemporal de los eventos sísmicos superficiales (profundidad menor o igual a 30 km) ocurridos en Colombia desde 2009 hasta 2024 mediante la aplicación de diferentes enfoques de clustering. La base de datos fue proporcionada por el Servicio Geológico Colombiano y el análisis se realiza utilizando características como la magnitud, profundidad, latitud y longitud de cada evento. También, se implementó un riguroso filtro basado en los errores de localización para asegurar la precisión y fiabilidad de los resultados obtenidos.

Además, se pretende contrastar la distribución de los sismos y las estructuras geológicas subyacentes, para identificar posibles patrones emergentes y precursores sísmicos. Este enfoque permite no solo entender mejor la dinámica sísmica regional, sino también aportar información valiosa para la toma de decisiones en políticas de mitigación del riesgo sísmico.

INTRODUCCIÓN

Colombia es un país sísmicamente activo, ya que se encuentra ubicado en la esquina noroccidental de Suramérica, donde convergen varias placas tectónicas (Bird, 2003). Estos eventos sísmicos son un desafío constante para la seguridad pública y la infraestructura, por lo que la gestión eficaz del riesgo sísmico requiere de una comprensión profunda y detallada de los patrones de actividad sísmica, planteando la necesidad de métodos avanzados para analizar y agrupar datos sísmicos. Este desafío se convierte en un problema clave para las autoridades de gestión de desastres y planificación urbana, quienes son los clientes potenciales de este análisis.

El contexto organizacional en el que surge este problema incluye agencias gubernamentales como el Servicio Geológico Colombiano (SGC), así como instituciones académicas y organizaciones dedicadas a la gestión de emergencias como la Unidad Nacional para la Gestión del Riesgo de Desastres (UNGRD). Estas entidades buscan mejorar sus capacidades de previsión y respuesta ante eventos sísmicos, lo que subraya la necesidad de herramientas analíticas más sofisticadas.

Este estudio se enfoca en aplicar técnicas de aprendizaje no supervisado para resolver el problema de la agrupación espaciotemporal de eventos sísmicos en Colombia entre 2009 y 2024. La pregunta principal que se aborda es cómo identificar patrones emergentes en la distribución de los sismos que puedan ayudar en la modelización de la actividad sísmica y en la previsión de futuros terremotos.

El aprendizaje no supervisado es adecuado para este tipo de análisis por la naturaleza compleja y no etiquetada de los datos sísmicos. Este estudio pertenece al área de **clustering**, una rama del aprendizaje no supervisado que se enfoca en agrupar datos en subconjuntos homogéneos

basados en similitudes inherentes. Al aplicar técnicas de clustering, se espera descubrir estructuras y patrones significativos en los datos sísmicos, aportando así soluciones prácticas y valiosas para la gestión del riesgo sísmico en Colombia.

REVISIÓN PRELIMINAR DE LA LITERATURA

El análisis de clustering aplicado a datos sísmicos ha experimentado un notable avance en la última década, con el objetivo de mejorar la comprensión de los patrones espaciotemporales de los terremotos y contribuir a una evaluación más precisa del riesgo sísmico.

Una aplicación común de clustering en sismología es la identificación y caracterización de "hotspots" o zonas con alta actividad sísmica. Muchos de estos estudios utilizan el algortimo K-Means, dada su simplicidad (p. ej. Kertanah *et al.*, 2022; Mato y Toulkeridis, 2017). Por ejemplo, Novianti *et al.* (2017) realizaron un estudio en una provincia de Bengkulu (Indonesia), donde se encuentran dos fallas geológicas activas, ellos agruparon epicentros de terremotos utilizando la latitud, longitud y magnitud de los eventos sísmicos a través del método de clustering K-Means con distancia euclidiana, donde el número óptimo de clusters se determinó utilizando el índice de Krzanowski y Lai (KL).

En un estudio más reciente y avanzado, Yuan (2021) propone incluso un algoritmo mejorado de clustering K-means para catálogos globales de terremotos. Este nuevo enfoque emplea una distancia espacio–tiempo–magnitud (STM) y luego propone utilizar una distancia STM máxima–mínima para la selección de los centros de los clusters iniciales. Además, se aplican el error cuadrático medio, el índice de Davies–Bouldin, el índice de Calinski–Harabasz y el coeficiente de Silhouette para determinar el número de clusters. Posteriormente, se presenta un modelo de predicción sísmica basado en el resultado del clustering combinado con una red neuronal artificial.

Aunque en diversas partes del mundo se han realizado numerosos estudios, en Colombia son pocos los trabajos en este tema, y la mayoría se han concentrado en la región occidental del país. Entre estos se encuentra el trabajo de Benítez *et al.* (2013), el cual utiliza una base de datos sísmica del suroeste de Colombia y aplica una técnica de clustering con partición difusa, desarrollado por Gath y Geva (1989) a partir del K-Means tradicional, donde se utiliza una distancia exponencial basada en la teoría de la máxima verosimilitud (Abonyi *et al.*, 2002). Por otra parte, Gracia (2017) realizó un estudio cerca al Pacífico Colombiano con 134 eventos sísmicos, empleando K-Means y analizando semivariogramas para estudiar el grado de correlación entre los eventos.

En este trabajo, al igual que en los estudios previamente presentados, se utilizará el enfoque de K-means. No obstante, también se explorarán otras metodologías como K-medoides y DBSCAN, que pueden ofrecer mayor robustez en la formación de los clusters. Además, a diferencia de algunos estudios donde no se realizó un preprocesamiento o filtrado de datos para no incluir eventos sísmicos con errores de localización, en este trabajo se aplicarán estos procedimientos para aumentar la confiabilidad de los resultados, y durante la selección también se tendrán en cuenta conocimientos técnicos sobre la evolución de la red de estaciones sismológicas, con la cual se recopila la información, ya que puede impactar en la precisión de los datos. Por otra parte, este estudio no sólo abordará el occidente de Colombia, sino todo el territorio.

DESCRIPCIÓN DE LOS DATOS

Los datos fueron proporcionados por el Servicio Geológico Colombiano y pueden ser obtenidos mediante el siguiente link: http://bdrsnc.sgc.gov.co/paginas1/catalogo/index.php.

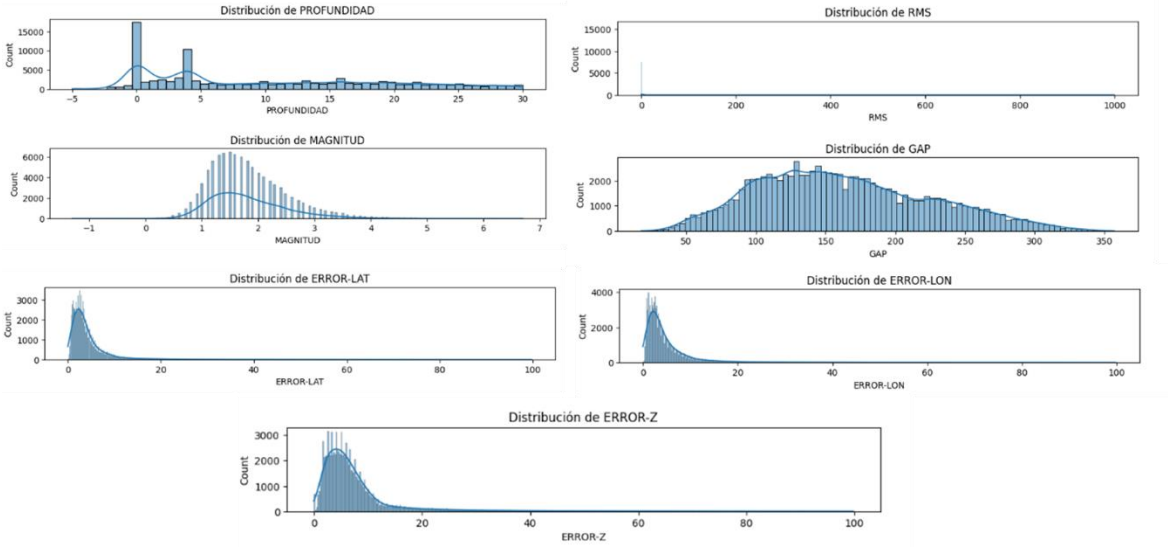
La base de datos contiene las características de 96177 eventos sísmicos superficiales (menores a 30 km de profundidad) ocurridos entre el 3 de junio 1993 y 30 de junio de 2024, ya que este tipo de sismicidad se relaciona con la distribución de estructuras geológicas como fallas activas, y tiene mayor potencialidad para causar daños significativos. En la Tabla 1 se describen estas variables.

Tabla 1. Descripción de variables de la base de datos

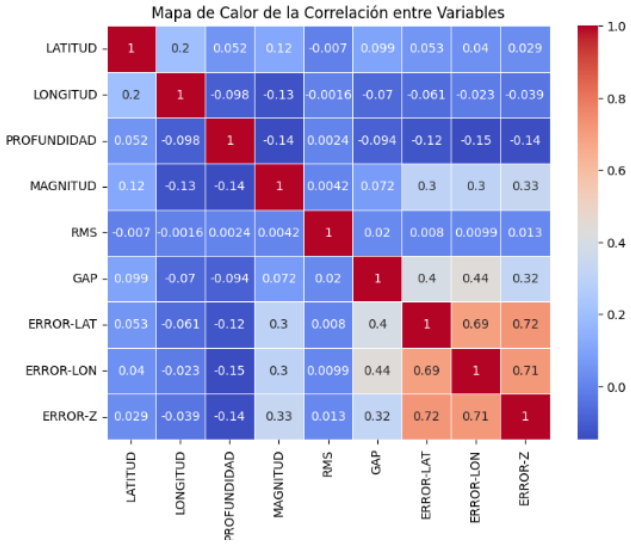
Variable	Tipo de variable	Descripción
Fecha	Datetime	La fecha en la que ocurrió el evento sísmico
Hora UTC	Datetime	La hora exacta del evento sísmico en Tiempo Universal Coordinado (UTC).
Latitud	Numérica (Float)	La latitud geográfica del epicentro del sismo
Longitud	Numérica (Float)	La longitud geográfica del epicentro del sismo
Profundidad	Numérica (Float)	La profundidad a la que ocurrió el sismo bajo la superficie terrestre, medida en km
Epicentro	String	El nombre del centro poblado más cercano al epicentro del sismo.
Magnitud	Numérica (Float)	La magnitud del evento sísmico. Este valor cuantifica la energía liberada
RMS	Numérica (Float)	Es la medida del error, comparando la diferencia promedio entre el tiempo de arribo teórico y el tiempo de arribo observado en segundos, utilizando las lecturas de los sismogramas. Los valores menores reflejan buenas localizaciones.
GAP	Numérica (Float)	El ángulo en grados como medida de la distribución angular entre el epicentro del sismo y las estaciones sísmicas que registran el evento. Entre mayor, hubo una menor cobertura de la red para registrar el sismo.
Error-Lat	Numérica (Float)	El margen de error en la estimación de la latitud del epicentro, medido en km
Error-Lon	Numérica (Float)	El margen de error en la estimación de la longitud del epicentro, medido en km
Error-Z	Numérica (Float)	El margen de error en la estimación de la profundidad del epicentro, medido en km

En las imágenes 1 y 3 se muestran las estadísticas descriptivas y los histogramas más destacados. Se observa que la profundidad media está alrededor de 10 km y la desviación estándar alta indica variabilidad en la profundidad de los sismos, lo que se puede corroborar en el histograma, sin embargo, muchos se concentran en los primeros 5 km. Respecto a la magnitud, se evidencia que la mayoría de los sismos presentan magnitudes bajas (alrededor de 1.8), aunque se han registrado sismos de hasta 6.7. El RMS exhibe una gran variabilidad, con algunos valores extremadamente altos de hasta 999 que podrían ser errores del software de localización de los eventos. Por lo que estos valores atípicos refuerzan la importancia de incluir en la metodología una etapa de preprocesamiento o filtrado de datos, puesto que el 75% presenta valores menores o iguales a 0.7. Lo mismo sucede con errores en latitud, longitud y profundidad, donde se observan valores atípicos de hasta 100 km.

	FECHA	LATITUD	LONGITUD	PROFUNDIDAD	MAGNITUD	RMS	GAP	ERROR-LAT	ERROR-LON	ERROR-Z
count	96177	96177.000000	96177.000000	96177.000000	96177.000000	96177.000000	96177.000000	96177.000000	96177.000000	96177.000000
mean	2016-03-17 19:53:52.128263424	5.278412	-75.171188	9.872022	1.818032	1.618084	163.435997	5.094376	5.796056	8.311757
min	1993-06-03 00:00:00	-3.483000	-85.633000	-4.971000	-1.300000	0.000000	17.800000	0.000000	0.000000	0.000000
25%	2013-04-30 00:00:00	3.434000	-76.255000	1.800000	1.300000	0.400000	116.000000	1.838478	1.848134	3.400000
50%	2018-08-11 00:00:00	5.128000	-75.124000	7.100000	1.700000	0.500000	156.000000	3.040559	3.200000	5.579915
75%	2021-04-06 00:00:00	6.922000	-74.093000	17.200000	2.200000	0.700000	206.000000	5.020458	5.900000	8.600000
max	2024-06-30 00:00:00	14.735000	-70.010000	30.000000	6.700000	999.000000	357.000000	99.800000	99.800000	99.800000
std	NaN	2.413719	1.757632	9.078527	0.705622	32.339725	62.383332	8.113271	9.151191	10.571160

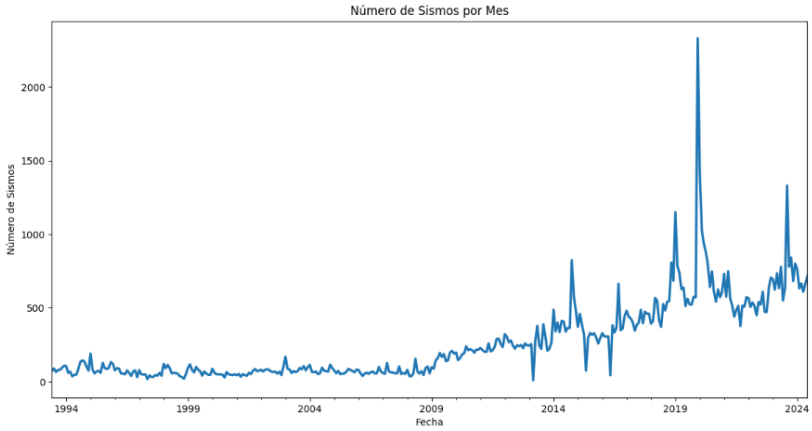


En el mapa de calor de correlación, se evidencia que en gran parte de las variables no existe una relación lineal importante que pueda generar problemas de multicolinealidad. A excepción de los errores de localización, donde se obtienen valores alrededor de 0,7; sin embargo, estas variables junto con el RMS y GAP solo se utilizarán para la etapa de preprocesamiento, puesto que no proporcionan información intrínseca de los eventos sísmicos.



Por otra parte, se graficó el registro de sismicidad en el tiempo y se observa un cambio considerable desde 2009, el cual se relaciona con la evolución de la red de estaciones sismológicas, dado que en ese año se aumentó considerablemente el número de estaciones. Por lo tanto, para evaluar datos obtenidos en condiciones similares, se tomarán aquellos sismos ocurridos entre 2009 y 2024. Aunque también se observan picos importantes posterior al 2009, estos ya se relacionan con la actividad sísmica del país, por ejemplo, a finales de 2019 se

observa un aumento importante asociado a las réplicas del sismo de magnitud 6.0 ocurrido en Mesetas (Meta).



PROPUESTA METODOLÓGICA

El objetivo de este proyecto es abordar la agrupación espaciotemporal de sismos en Colombia mediante la implementación de técnicas de clustering. Se ha seleccionado el algoritmo K-means como la técnica principal debido a su capacidad para dividir eficientemente un conjunto de datos en grupos (clusters) que minimizan la varianza dentro de cada grupo. Este método es particularmente adecuado para este tipo de análisis por su simplicidad y eficiencia computacional, lo que lo hace ideal para grandes volúmenes de datos como los de sismicidad.

Para este análisis, se ha decidido no aplicar Análisis de Componentes Principales (PCA) ni Descomposición en Valores Singulares para reducir dimensiones, debido a la cantidad limitada de variables (magnitud, profundidad, latitud y longitud). La conservación de todas las dimensiones originales permitirá una evaluación completa de las características sísmicas sin pérdida de información relevante.

Reconociendo las limitaciones inherentes de K-means, como su sensibilidad a outliers y la necesidad de predefinir el número de clusters, se complementará este análisis con K-medoides. Este algoritmo utiliza la distancia de Manhattan para seleccionar medoids, lo que lo hace menos susceptible a valores atípicos y más robusto en la representación de clusters no esféricos. Ambos algoritmos serán implementados utilizando la librería scikit-learn en Python, aprovechando sus optimizaciones para el manejo eficiente de grandes volúmenes de datos.

Además de K-means y K-medoides, se consideraron otros algoritmos como DBSCAN y Hierarchical Clustering, pero se decidió no incluirlos en esta fase inicial debido a la naturaleza específica de los datos y a las limitaciones de tiempo. Sin embargo, estos algoritmos podrían ser explorados en etapas posteriores del proyecto si se considera necesario.

Para evaluar la calidad de los clusters, se utilizarán métricas como el índice de Silhouette Estas medidas permitirán una validación rigurosa de la agrupación obtenida, asegurando que los resultados sean interpretables y de utilidad para la gestión del riesgo sísmico en Colombia.

BIBLIOGRAFÍA

Abonyi, J., Babuska, R., & Szeifert, F. (2002). Modified Gath-Geva fuzzy clustering for identification of Takagi-Sugeno fuzzy models. *IEEE Transactions on Systems, Man, and*

Cybernetics, Part B (Cybernetics), *32*(5), 612–621. https://doi.org/10.1109/TSMCB.2002.1033180

Benítez, H. D., Flórez, J. F., Duque, D. P., Benavides, A., Lucía Baquero, O., y Quintero, J. (2013). Spatial pattern recognition of seismic events in South West Colombia. *Computers & Geosciences*, *59*, 60–77. https://doi.org/https://doi.org/10.1016/j.cageo.2013.04.028

Bird, P. (2003). An updated digital model of plate boundaries. *Geochemistry, Geophysics, Geosystems*, *4*(3). https://doi.org/10.1029/2001GC000252

Gath, I., y Geva, A. B. (1989). Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(7), 773–780. https://doi.org/10.1109/34.192473

Gracia, M. D. (2017). *Seismotectonic characterization of the Colombian Pacific region: Identification of tectonic patterns through geostatistical analysis* (Undergraduate thesis). Universidad de los Andes, School of Sciences, Faculty of Geosciences.

Kertanah, K., Rahadi, I., Aryani Novianti, B., Syahidi, K., Sapiruddin, S., Mandala Putra, H., Gazali, M., Haiban Hirzi, R., y Sabar, S. (2022). APPLYING K-MEANS ALGORITHM FOR CLUSTERING ANALYSIS EARTHQUAKES DATA IN WEST NUSA TENGGARA PROVINCE. *Indonesian Physical Review*, *5*(3), 197–207. https://doi.org/10.29303/ipr.v5i3.148

Mato, F., y Toulkeridis, T. (2017). An unsupervised K-means based clustering method for geophysical post-earthquake diagnosis. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1–8. https://doi.org/10.1109/SSCI.2017.8285216

Novianti, P., Setyorini, D., y Rafflesia, U. (2017). K-Means cluster analysis in earthquake epicenter clustering. *International Journal of Advances in Intelligent Informatics*, *3*(2), 81. https://doi.org/10.26555/ijain.v3i2.100

Yuan, R. (2021). An improved K-means clustering algorithm for global earthquake catalogs and earthquake magnitude prediction. *Journal of Seismology*, *25*(3), 1005–1020. https://doi.org/10.1007/s10950-021-09999-8