Identificación de zonas sísmicas importantes a través de clusterización y puntos calientes

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. Esto subraya la importancia de analizar los catálogos de sismicidad para ofrecer valiosas aproximaciones para modelizar la actividad sísmica.

En este trabajo se abordó el desafío de la agrupación 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 y obtención de puntos calientes. La base de datos fue proporcionada por el Servicio Geológico Colombiano y se empleó características como magnitud y profundidad de cada evento.

Lo mejores resultados fueron obtenidos con Kernel Density Estimation (KDE) y DBSCAN puesto que identifica patrones importantes de sismicidad que coinciden con la distribución de fallas activas, las cuales pueden ser generadoras de sismos. Se identificaron zonas en los departamentos del Meta, Huila, Chocó, Norte de Santander y Santander, lo que resulta valioso para la implementación de estrategias de densificación de estaciones sismológicas, así como para mejorar los planes de gestión del riesgo y ordenamiento territorial de dichas regiones.

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 para la seguridad pública y la infraestructura, por lo que la gestión 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, especialmente en zonas vulnerables, 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 de eventos sísmicos en Colombia entre 2009 y 2024. La pregunta principal abordada es cómo identificar patrones importantes y emergentes de actividad sísmica que permitan fortalecer las estrategias de gestión del riesgo sísmico y de ordenamiento territorial.

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 y puntos calientes, una rama del aprendizaje no supervisado que se enfoca en agrupar datos en subconjuntos homogéneos basados en similitudes inherentes. Al aplicar estas técnicas, 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.

MATERIALES Y MÉTODOS

Los datos utilizados en este análisis fueron proporcionados por el Servicio Geológico Colombiano y pueden ser consultados en el siguiente enlace: http://bdrsnc.sgc.gov.co/paginas1/catalogo/index.php.

La base de datos incluye las características de 96177 eventos sísmicos superficiales (menores a 30 km de profundidad) registrados entre el 3 de junio 1993 y 30 de junio de 2024. Se seleccionaron estos sismos superficiales debido a su relación con estructuras geológicas como fallas activas, las cuales tienen un mayor potencial para causar daños significativos. En la Tabla 1, se describen las variables incluidas en la base de datos.

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

La Figura 1 muestra las estadísticas descriptivas e histogramas. El análisis revela que la profundidad media de los sismos es de 10 km, con alta variabilidad, concentrándose muchos eventos en los primeros 5 km. La magnitud promedio es de 1.8, con algunos eventos de hasta 6.7. El RMS muestra gran dispersión, con valores extremos que alcanzan 999, indicando posibles errores en la localización, lo que sugiere la necesidad de filtrar outliers. Además, se detectaron errores de hasta 100 km en las coordenadas, subrayando la importancia del preprocesamiento.

El mapa de calor de correlación indica que no hay relaciones lineales significativas entre la mayoría de las variables, lo que evita problemas de multicolinealidad. Sin embargo, hay correlaciones moderadas (0.7) entre los errores de localización, que, junto con el RMS y el GAP, se usaron solo en el preprocesamiento (filtrado) porque no proporcionaba información intrínseca. La figura también muestra un aumento en la sismicidad desde 2009, coincidiendo con la expansión de la red sismológica en Colombia, por lo que se filtraron los datos desde este año y se calcularon los días transcurridos desde el 01-01-2019 para identificar posibles patrones temporales de sismicidad.

En este análisis, se optó por no usar Análisis de Componentes Principales (PCA) ni Descomposición en Valores Singulares (SVD) debido a la cantidad y naturaleza de las variables involucradas. En su lugar, se aplicaron varias técnicas de clustering para identificar zonas de sismicidad, como k-medias, k-medoides, jerárquico aglomerativo, DBSCAN, KDE y KDE bivariado.

	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

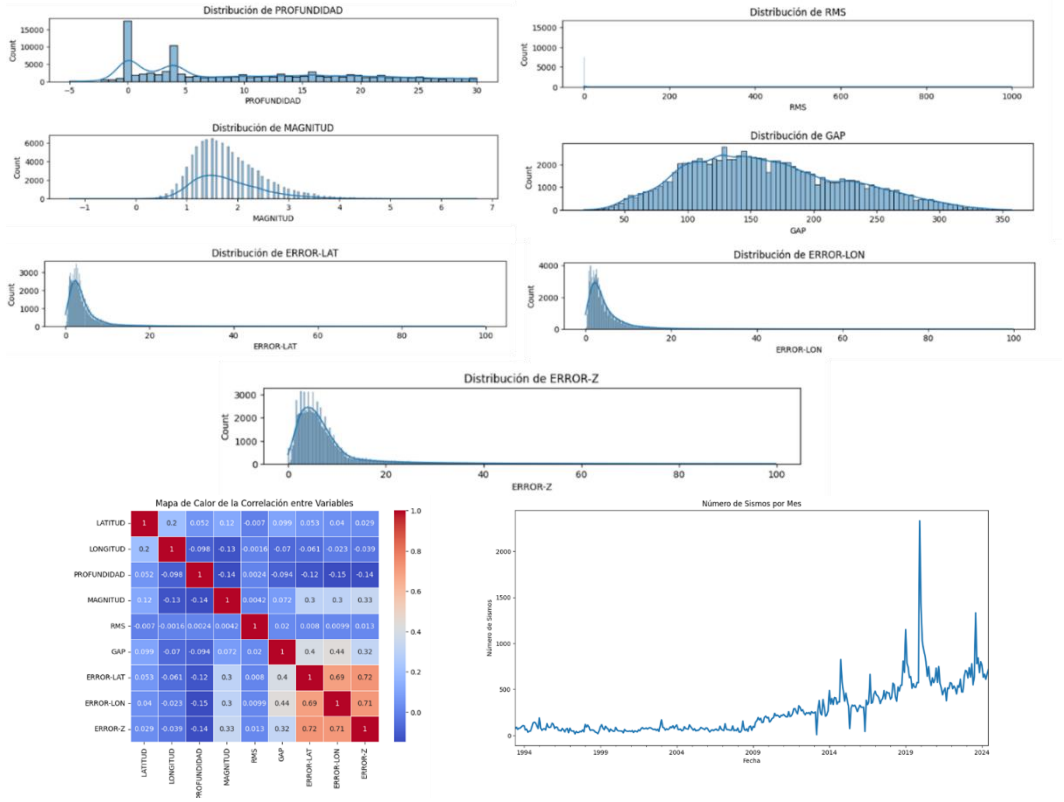

Figura 1. Estadísticas descriptivas e histogramas para las variables Profundidad, RMS, Magnitud, GAP, Error-Lat, Error-Long y Error-Z. Mapa de correlación y gráficos de tiempo vs número de sismos registrados

1. **K-medias y k-medoides**: Para estos métodos, se determinó el número óptimo de clústeres mediante el método del codo y el Silhouette Score. K-medias fue elegido por su eficiencia en grandes conjuntos de datos, mientras que K-Medoides es más robusto ante valores atípicos.
2. **Jerárquico aglomerativo**: Se utilizó este método para analizar la estructura de los datos sin necesidad de definir un número fijo de clústeres a priori. Se generó un dendrograma que permitió visualizar la fusión de los clústeres en diferentes niveles, ayudando a identificar el número óptimo de agrupaciones basado en la estructura jerárquica de los datos. Sin embargo, su escalabilidad es limitada para grandes datasets.
3. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: Este método fue utilizado para identificar clústeres de forma basada en densidad, lo que lo hace eficiente en la detección de patrones en zonas con ruido (outliers) y formas irregulares. Para seleccionar los mejores hiperparámetros, primero se optimizó la distancia epsilon mediante el método del codo y luego se ajustó el número mínimo de puntos en cada clúster.
4. **KDE y KDE bivariado**: Se utilizaron técnicas de estimación de densidad de kernel (KDE) para detectar regiones con alta densidad de eventos sísmicos. En el KDE bivariado, se consideraron simultáneamente dos dimensiones, proporcionando una visión más precisa de la distribución de sismos en el espacio.

Para comparar los resultados de clustering, se empleó el Índice de Davies-Bouldin, una métrica que evalúa la dispersión interna y la separación entre clústeres. No se usó el Silhouette Score para esta comparación final, dado que ya había sido utilizado para definir el número óptimo de clústeres, evitando así un sesgo en la evaluación. Los resultados fueron visualizados mediante boxplots, lo que permitió una fácil identificación de las características distintivas de cada clúster, y luego se contrastaron con información técnica existente para asegurar la coherencia de los enfoques empleados.

RESULTADOS Y DISCUSIÓN

Con los métodos de clustering se buscó identificar grupos con magnitudes altas, profundidades cercanas a 0 (muy superficiales) y si es posible el patrón temporal (sismos recientes o antiguos), ya que aquellos con estas características son los que presentan mayor potencial para causar daños significativos.

En el caso de K-medias y K-medoides, inicialmente se definió un número óptimo de clústeres, pero al analizar las características de cada grupo, no se obtuvo información significativa. Finalmente, se estableció el número óptimo en 21 clústeres, basado en los cambios observados en la gráfica del codo y el Silhouette Score. Se destacaron algunos clústeres con patrones significativos, especialmente en la Cordillera Oriental y cerca de la frontera con Panamá. Para el Clustering Jerárquico Aglomerativo, se utilizó la distancia euclidiana y el método de Ward. Mediante el análisis del dendrograma, con una distancia óptima de corte en 35, se identificaron 19 clústeres. Sin embargo, la mayoría de estos no presentaron variaciones relevantes, limitando su utilidad. En cuanto a DBSCAN, se realizaron 18 pruebas para afinar los hiperparámetros, definiendo los valores óptimos de eps en 0.89 y min_samples en 4, lo que resultó en 31 clústeres. Aunque muchos eventos se agruparon en el clúster 0 o se clasificaron como ruido, los otros presentaron características distintivas en magnitud, profundidad y tiempo transcurrido, sugiriendo patrones importantes pese a su baja densidad. Algunos de los clusters más importantes se encuentran cerca de Villavicencio (Meta), Pitalito (Huila), Ipiales (Nariño) y Murindó(Chocó).

Al comparar los resultados, se confirma el bajo rendimiento del clustering aglomerativo, posiblemente debido a su ineficiencia en grandes conjuntos de datos, obteniendo el mayor valor de Davies-Bouldin. DBSCAN mostró una mejor diferenciación, pero fue penalizado por sus clústeres dispersos y de gran tamaño. Los mejores resultados, según el índice, fueron de K-medias y K-medoides, aunque con la limitación de que el número óptimo de clústeres depende de la selección inicial de centroides. Los resultados se presentan en la

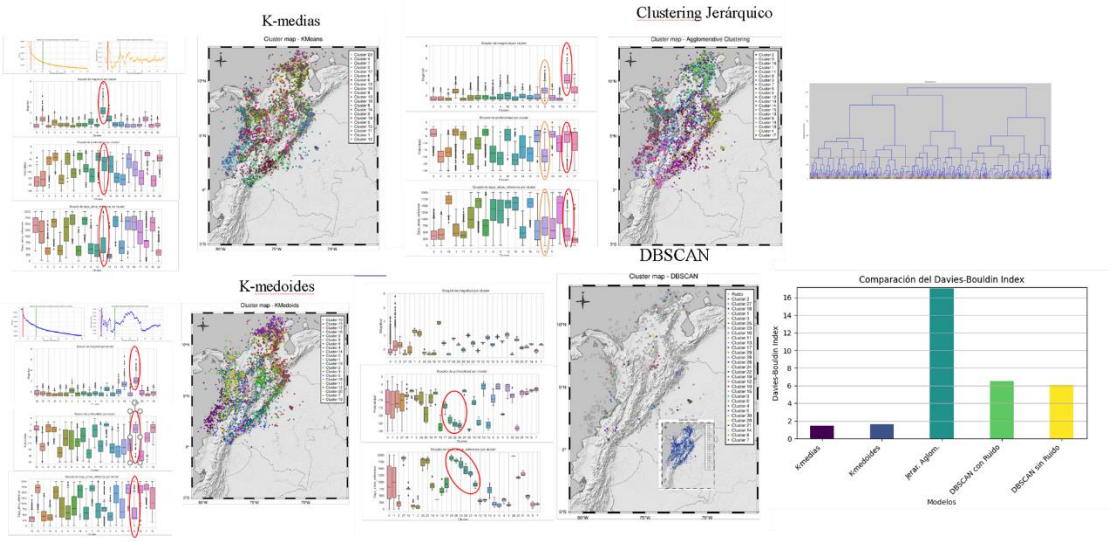

Figura 2. Resultados K-medias, Clustering Jerárquico Algomerativo, K-medoides, DBSCAN y comparación con el indice Davies-Bouldin

En los resultados obtenidos con KDE, se identificaron zonas de puntos calientes que coinciden con las detectadas por DBSCAN, como se muestra en la Figura 3. Al contrastar estos resultados con información geológica, se observó que DBSCAN, KDE normal y KDE bivariado ofrecen la mejor aproximación, ya que los patrones identificados coinciden en gran parte con la distribución de fallas activas, muchas de las cuales son generadoras de sismos, aunque no siempre se conoce su naturaleza exacta. Un hallazgo clave con DBSCAN fue la identificación de un patrón en Antioquia, lo que sugiere la necesidad de una mayor densificación de estaciones sismológicas por parte del SGC en esa región. Esto refuerza la capacidad de DBSCAN para detectar patrones, especialmente en zonas con sismos superficiales y de magnitud considerable, dado que el algoritmo maneja bien los datos atípicos (Figura 3).

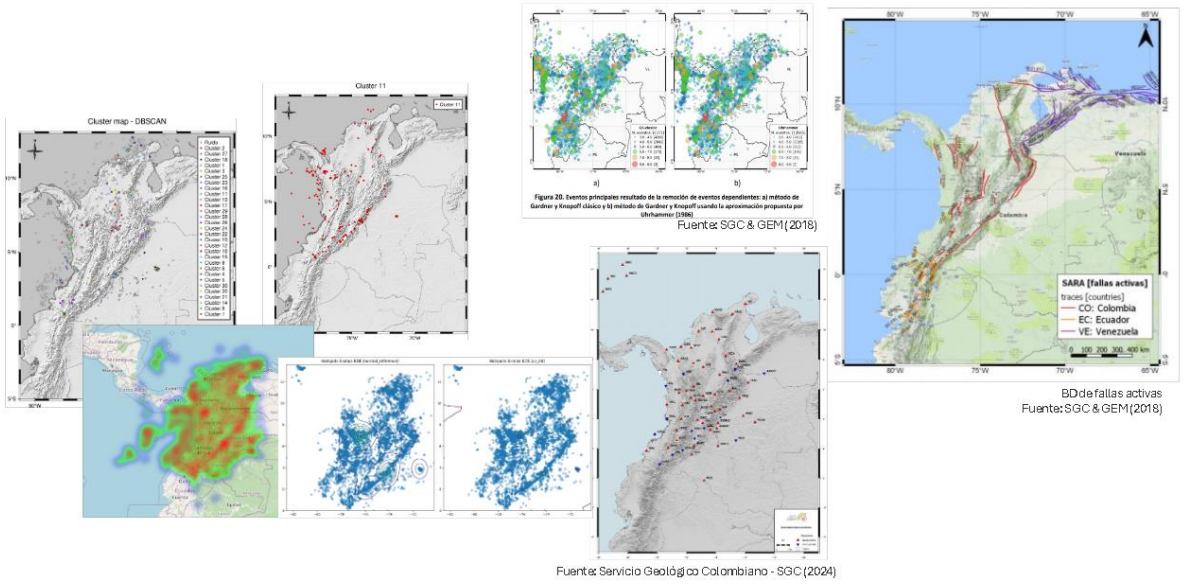

Figura 3. Resultados DBSCAN, Cluster destacado de K-medias, KDE, KDE Bivariado. A la izquierda están mapas que permitieron contrastar los resultados. Se encuentra el mapa de estaciones sismológicas, sismicidad de Colombia por magnitud y distribución de fallas activas.

Estos resultados no solo permiten mejorar los planes de gestión del riesgo y el ordenamiento territorial, sino que también son fundamentales para planificar la densificación de estaciones sismológicas en zonas críticas. Además, resaltan áreas de interés para futuras investigaciones sobre fallas activas potencialmente generadoras de sismos. Para estudios futuros, sería útil aplicar estas técnicas incluyendo sismicidad intermedia y profunda, así como ajustar los parámetros de filtrado de datos. La combinación de múltiples técnicas también podría mejorar los resultados obtenidos.

CONCLUSIONES

- Se aplicaron los métodos de k-medias, k-medoides, jerárquico aglomerativo, DBSCAN, KDE y KDE bivariado para identificar zonas de sismicidad.
- El método jerárquico aglomerativo mostró el peor desempeño, posiblemente debido a su ineficiencia para grandes conjuntos de datos. En contraste, los métodos de k-medias y k-medoides demostraron ser más

efectivos según el Índice de Davies-Bouldin (IDB). Aunque DBSCAN no obtuvo el mejor desempeño en términos del IDB, demostró ser eficiente en la identificación de áreas clave.

- Se identificaron zonas de alta sismicidad cerca de Mesetas y Villavicencio (Meta), Gigante y Pitalito (Huila), Sur de Nariño, Murindo, Río Sucio (Chocó), Cucuta y Pamplona (Norte de Santander) y Barrancabermeja (Santander) lo que resulta valioso para la implementación de estrategias de densificación de estaciones sismológicas en estas zonas, así como para mejorar los planes de gestión del riesgo y ordenamiento territorial. Esto permite identificar patrones de actividad sísmica y ajustar las políticas preventivas en consecuencia.
- En este análisis, solo se consideró la sismicidad superficial (menos de 30 km de profundidad). No obstante, sería útil explorar para futuras investigaciones estas técnicas para incluir sismicidad intermedia y profunda, así como ajustar los parámetros de filtrado de datos. Además, la combinación de varias técnicas podría mejorar los resultados obtenidos.

BIBLIOGRAFÍA

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

SGC y GEM (2018). "Modelo Nacional de Amenaza sísmica de Colombia. Servicio Geológico Colombiano (SGC) – Grupo de Amenaza Sísmica. Fundación Global Earthquake Model (GEM). 196 pp.