Algoritmo *k-means* para agrupamiento de segmentos en sismogramas mediante descriptores: 2 casos de estudio

Objetivos:

1. Segmentación de un sismograma mediante algoritmos de aprendizaje no supervisado (*unsupervised machine learning*), en este caso *K-MEANS*, haciendo uso de descriptores estadísticos.

2. Determinar posibles cambios en el régimen de un sistema mediante la evolución temporal de las "familias" o agrupaciones determinadas por el algoritmo.

3. Determinar las características más importantes de cada "familia" o agrupación, basadas en el análisis de una señal guía o "madre" única de cada "familia".

Introducción:

El monitoreo y análisis de actividad sísmica en áreas volcánicas (Developments in Volcanology, 2003) y en zonas de carácter industrial, con antecedentes de sismicidad inducida (Zou, 2017), es de vital importancia para la gestión del riesgo y el desarrollo de sistemas de alerta temprana. El establecimiento de anomalías en una serie de tiempo de una variable física relacionada con un sistema puede estar vinculado con cambios en el régimen o en los procesos que allí ocurren. Un conocimiento más a fondo de estos cambios nos puede ayudar a pronosticar diferentes procesos en un mismo sistema. Sin embargo, debido al carácter de esta información, usualmente conlleva de gran dificultad para un observador humano revisar, analizar y correlacionar posibles patrones en una serie de tiempo de miles de datos. Es así, que este proyecto busca mediante la metodología planteada por (Watson, 2020), usar algoritmos de aprendizaje no supervisado (en este caso *k-means*) en descriptores de diferentes segmentos de un sismograma, en busca de patrones que indiquen cambios en el sistema físico. Para esto se creará una rutina en Python, que permita analizar 2 series de tiempo, vinculados a 2 procesos diferentes:

- Sismograma registrado en la estación sismológica GUY2C del SGC (perteneciente a la RSNC, CM), específicamente la localización 00 y el componente HHZ, entre el 7 y el 16 de noviembre del 2022, con intención de identificar cambios en el comportamiento de las "familias" o agrupaciones en el intervalo del 8 y 14 de noviembre del 2022 relacionada con una serie de boletines que registraban actividad sísmica y caída de cenizas en ese periodo de tiempo (SGC, 2022).

- Sismograma registrado en la estación sismológica LL8C del SGC (perteneciente a la RSNC, CM), específicamente la localización 00 y el componente HHZ, entre el 13 y el 19 de marzo del 2018, con intención de identificar cambios en el comportamiento de las "familias" o agrupaciones después del 14 de marzo del 2018 relacionada con la iniciación de procesos de fracturamiento hidráulico en perforación vertical en el campo minero La Loma, Cesar (https://www.elheraldo.co/economia/consejo-de-estado-ordena-suspender-produccion-de-gas-de-drummond-en-campo-la-loma-687184).

Cabe destacar que en ninguno de estas series de tiempo se contempla la posibilidad de que las señales o estímulos no correspondan al sistema o al proceso en sí, deben verse de manera más general como cambios donde el sensor este puesto. Aun así, cabe resaltar que su cercanía a estos hace que sus señales se vean de una u otra manera involucrada.

De igual manera, se puntualiza que el objetivo del proyecto es tener como resultado una evolución temporal de familias, mas no, un análisis exhaustivo de ellas. La creación de señales "madre", como segmentos identificadores de cada agrupación permite un análisis "somero" de estas.

Metodologia:

Partiendo de igual manera en los dos sismogramas se hacen los siguientes pasos, especialmente desde el punto 2-5 se refiere a la metodología planteada por (Watson, 2020), con ligeras variaciones:

1. Lectura y elección del componente HHZ de cada estación mediante la librería Obspy (Krischer et al., 2015). De igual manera se filtra el "ruido" de baja frecuencia, con un filtro "highpass" de 0.1 Hz. Si bien, puede haber ruido de mas altas frecuencias, la intención es que el algoritmo pueda detectarla. Sin embargo, tal como recomienda (Chong, 2021), el hecho de no hacer filtros lo suficientemente restrictivos puede dar lugar a resultados erróneo.
2. Se segmenta la señal en ventanas de tiempo de 60 segundos, con un porcentaje de traslape del 80%.
3. Para cada señal se le extraen una serie de descriptores (Watson, 2020), estos son
    - Desviación estándar en el domino del tiempo
    - Curtosis en el domino del tiempo
    - Asimetría en el domino del tiempo
    - Pico en el domino de frecuencia
    - Asimetría en el domino de frecuencia

    Estos descriptores, más allá de significar un fenómeno físico, resalta una semejanza estadística entre posibles agrupaciones.
4. Creación de una matriz en donde cada segmento tenga su representación en descriptores.
5. Utilización del algoritmo k-means. Este paso se divide en 2, uno de estimación del número de agrupaciones mediante diferentes métricas, y otro relacionado a la estimación para la serie de tiempo en sí.
    - El algoritmo *k-means*, es un algoritmo de agrupación iterativo, el cual debido a su naturaleza es categorizado en el ámbito de aprendizaje no supervisado. En este lo único que el usuario debe ingresar es el número de grupos, "clusters" o "familias". Para esto calcula el valor promedio de las distancias, dando un centroide inicial (un punto donde se calcularán las diferentes distancias, y el cual será el centro del área de cada familia dada las variables). Dado este resultado, se buscará una serie de centroides en donde la distancia se vea minimizada, hasta converger (Chong, 2021). El proceso de optimización, así como el algoritmo es representado en la siguiente ecuación:

Ecuación 1: k-means según (Chong, 2021)

$$\underset{S}{argmin} \frac{1}{n} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2$$

De esta manera es de vital importancia dar un valor para el numero de agrupamientos, usando en este caso 3 métricas para un rango de 2 a 19 agrupaciones (representados en la Figura 1y Figura 2 para el caso 1 y 2, respectivamente):

i. Codo (*Elbow* / Inercia): Uso de la distancia al cuadrado entre cada punto y su centroide. La suma de estos nos da la inercia. El punto de máxima curvatura corresponderá al valor óptimo de agrupaciones ((Chong, 2021).
ii. Davies Bouldin (*DB score*): Representa la similitud promedio de cada agrupación entre ellas. Entre menor es mejor, ya que permite discernir entre las agrupaciones (https://medium.com/mlearning-ai/deciding-number-of-clusters-using-gap-statistics-davies-bouldin-index-calinski-harabasz-index-2ce9acfb6118#:~:text=Davies%20Bouldin%20index%20is%20calculated,cluster%20distance%20to%20cluster%20size.)
iii. Silueta (*Silhouette*): Medida de similitud de un punto a en su propia agrupación (cohesión) comparando con otros (separación). Un valor alto indica que este punto esta emparejado de mejor manera con su propia agrupación que lo demás (https://es.wikipedia.org/wiki/Silhouette_(clustering)).

Nótese en la Figuras 1 y 2, la poca semejanza que tiene el codo con las demás métricas, y dado que estas otras indicas similitud, para términos de este proyecto indican un mejor numero de referencia. Sin embargo, en el caso 2 se usa también un k=7 ya que tiene relación con el *DB score*.

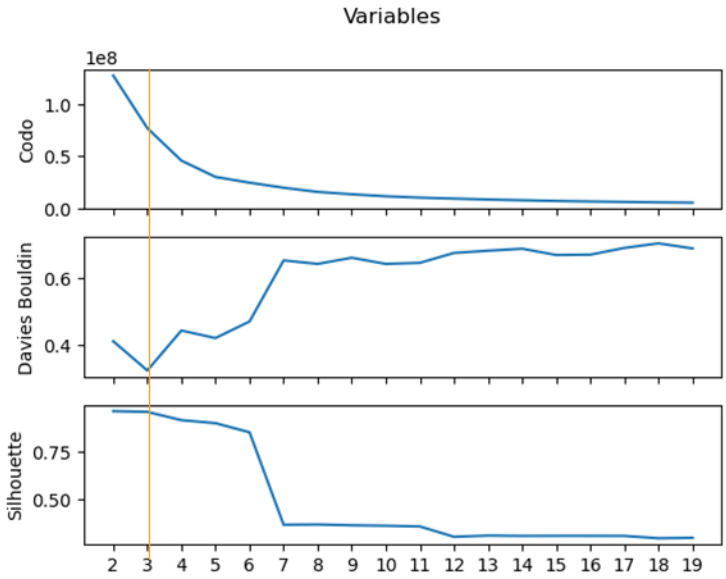

Figura 1: Métricas para el caso 1.

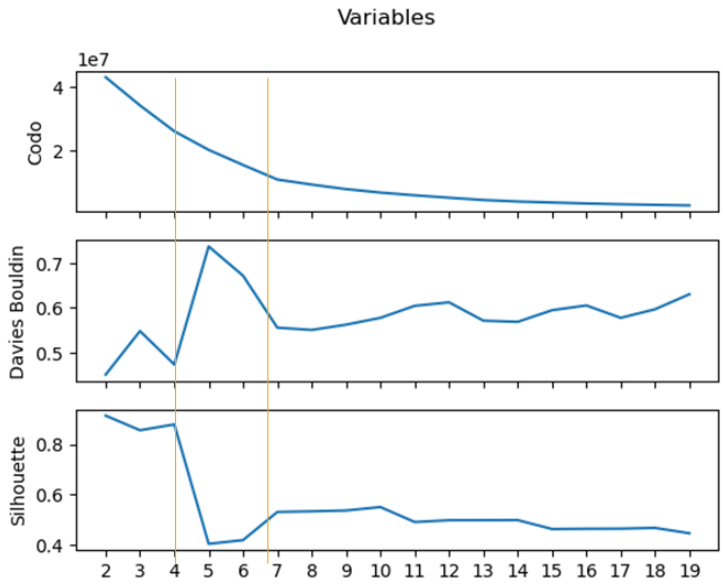

Figura 2: Métricas para el caso 2. Véase la posibilidad de un k en 4 y 7.

- Utilización del algoritmo

6. Uso de la distancia coseno/similitud (https://es.wikipedia.org/wiki/Similitud_coseno) para extraer las señales "madre", a partir del promedio de cada uno de los descriptores de la señal, en relación con cada segmento. La fórmula es representada en la siguiente ecuación:

Ecuación 2: Similitud dado la distancia de coseno

$$S = (a \cdot b)/(|a| * |b|)$$

Resultados:

1. Caso 1 – Nevado del Ruiz:

Dado el número de "familias" o agrupaciones como 3 (Figura 1), se establece la siguiente evolución temporal (Figura 3):

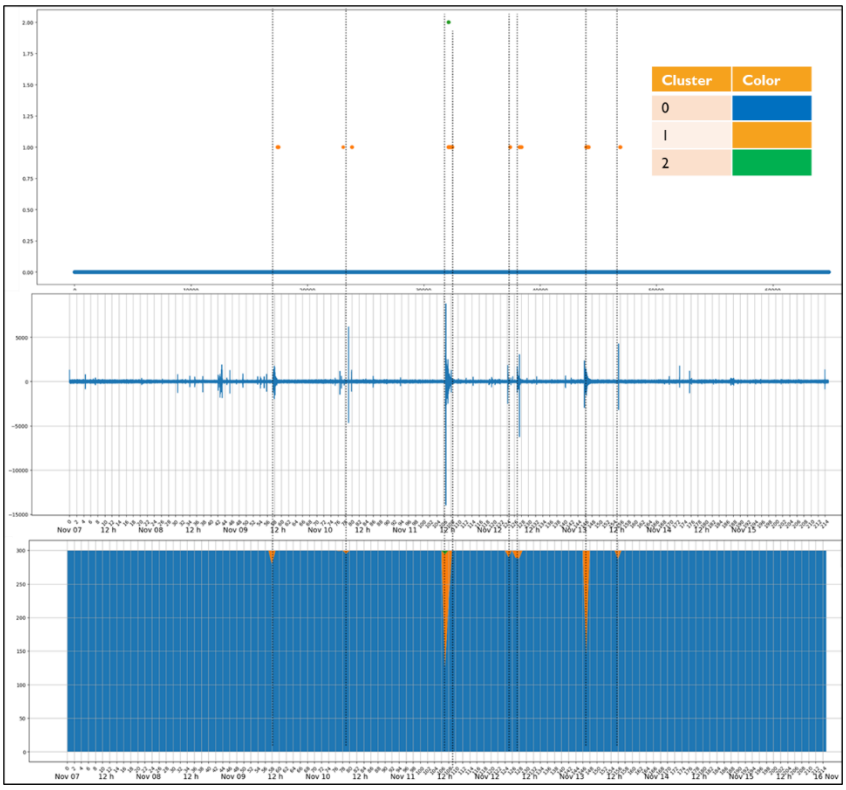

Figura 3: Evolución temporal de las "familias" determinadas a partir del algoritmo en el sismograma de la estación GUY2C, con correlaciones resaltadas a partir de líneas negras. k=3.

De igual manera dado la similaridad se establecen las señales "madre" de cada familia (Figura 4):

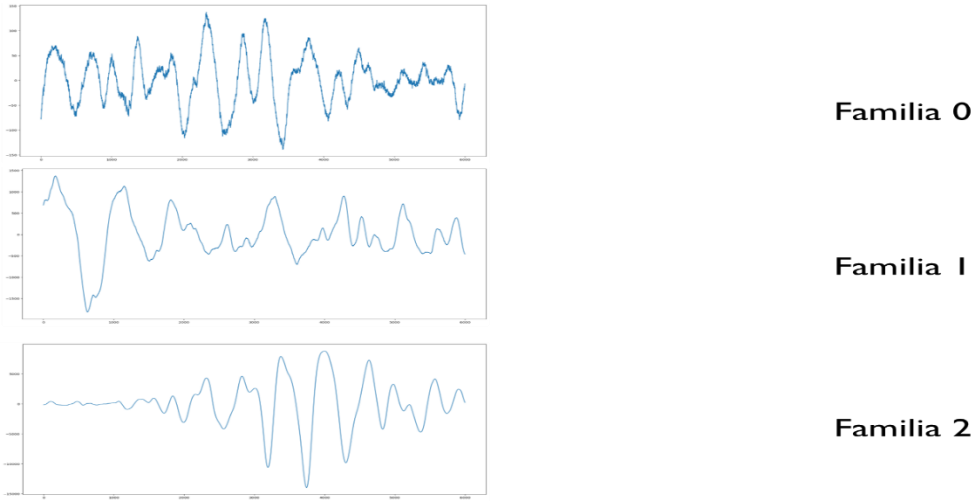

Figura 4: Señales "madre" de cada "familia" para el caso 1.

2. Caso 2 – La Loma, Cesar:

Dada la elección del número de familias (Figura 2) como 4 y 7 (Figura 5 y Figura 6, respectivamente), se establece la siguiente evolución temporal de estas:

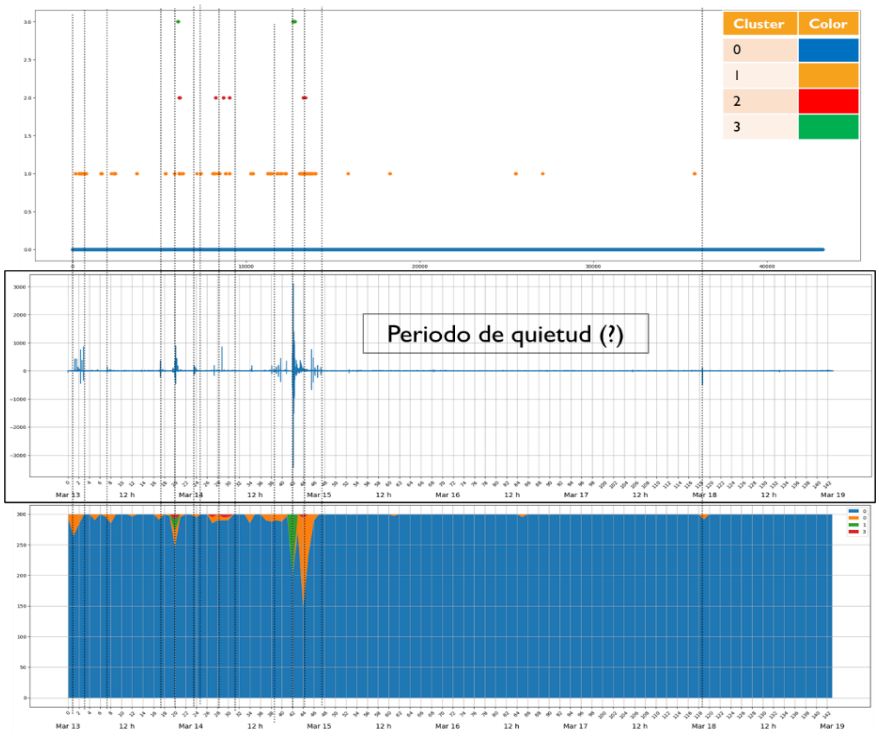

Figura 5: Evolución temporal de las "familias" determinadas a partir del algoritmo en el sismograma de la estación LL8C, con correlaciones resaltadas a partir de líneas negras. k=4.

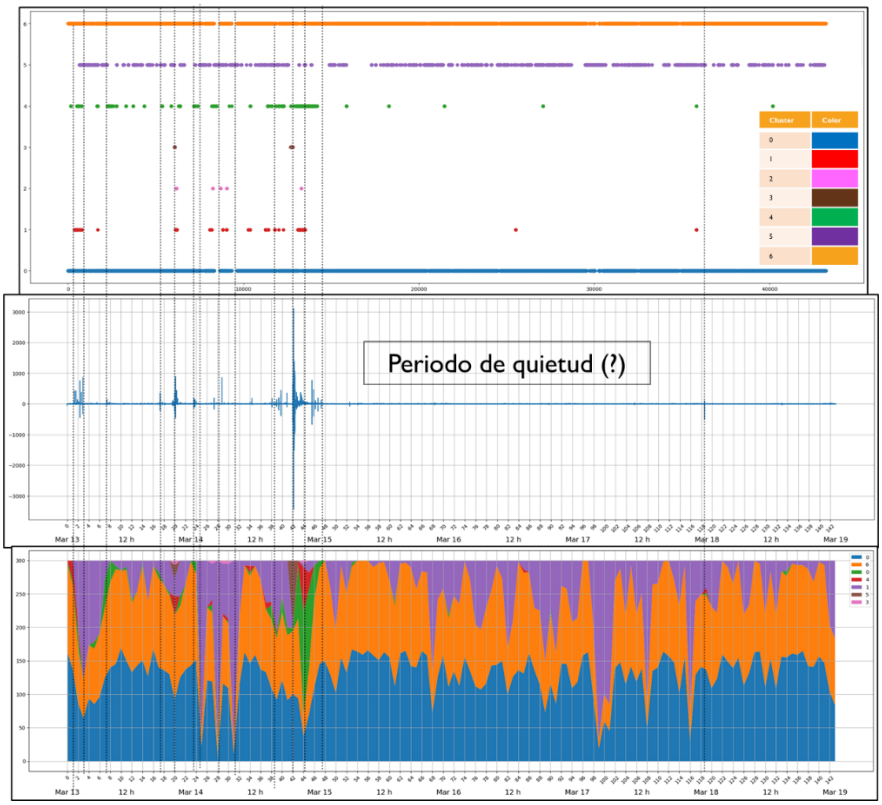

Figura 6: Evolución temporal de las "familias" determinadas a partir del algoritmo en el sismograma de la estación LL8C, con correlaciones resaltadas a partir de líneas negras. k=7.

De igual manera dado la similaridad se establecen las señales "madre" de cada familia (Figura 7), en este caso se usó la de 4 familias, ya que según las métricas es más representativo:

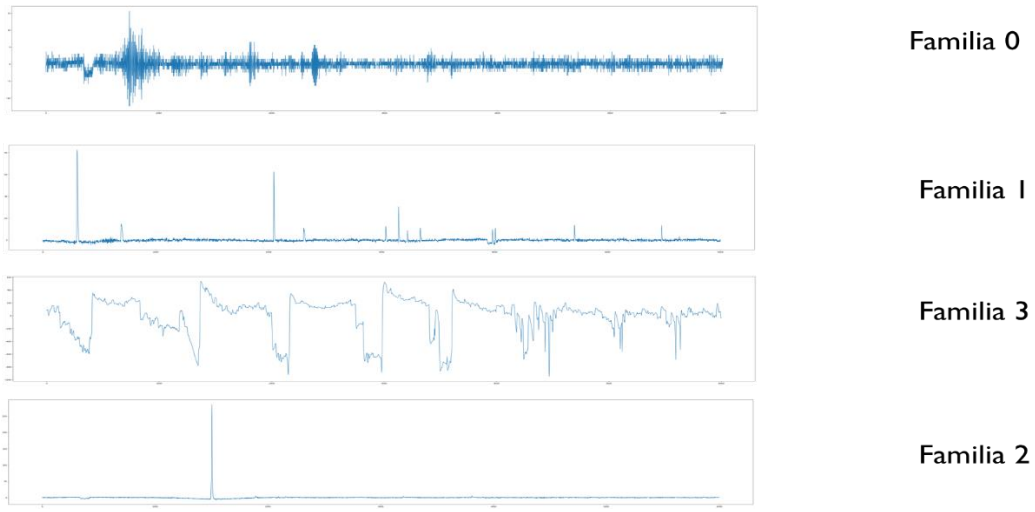

Figura 7: Señales "madre" de cada "familia" para el caso 2.

Análisis:

1. *Caso 1 – Nevado del Ruiz:*

Para el primer caso se establece una similitud muy buena entre eventos energéticos altos (amplitudes) con las familias 1 y 2, siendo mas visible en el caso de la 2. De esta manera, podríamos correlacionar la familia 1 como el comportamiento normal de la zona, pero no es muy bueno denotando cambios más pequeños. Es de esta manera que clasifica bien los eventos sísmicos relacionados a la actividad eruptiva del volcán entre las 12 h del 9 de noviembre y las 12 h del 14 de noviembre. En cuanto a las señales "madres" podemos concluir lo mismo, principalmente diferenciándolo por las diferencias en la magnitud de las amplitudes. Cabe destacar que la señal madre de la familia 2 resalta más la forma de un sismo convencional.

2. *Caso 2 – La Loma, Cesar:*

Para el segundo caso resalta muy bien las anomalías en la señal para k=4, siendo representados por las familias 1,2 y 3, mientras que la 0 podría representar el comportamiento normal de la zona. Para k=7, tenemos una variación bastante alta, principalmente resaltando las anomalías por las familias 1,2,3 y 4, mientras que un comportamiento normal con un grado de variación en el 0,5 y 6. Independientemente el k, los dos resaltan un periodo de quietud hacia el 15 de marzo, luego de un máximo de intensidad. En cuanto a la señales "madres" podemos concluir algo muy similar, dado por una inconsistencia en la familia 3, producto posiblemente de un error en el muestreo. Además, la familia 1 y 2 se comportan de manera impulsiva.

Conclusiones:

1. Se observan patrones los suficientemente diferentes para indicar anomalías a nivel sísmico.
2. Las señales cuentan con errores de carácter instrumental, los cuales pueden afectar la agrupación.
3. En algunos casos, hay discrepancia entre el método codo y los demás.

Se recomienda:

1. Añadir otros descriptores, por ejemplo, los contemplados en Cadena Ibarra, O. (2021)
2. Posibilidad de una ventana de tiempo más extensa, y una segmentación menor (mayor capacidad de cómputo).
3. Analizar de manera más profunda las señales "madre" de cada familia
4. Correlacionar con eventos más discretos (sismos, inyección de agua, flujos piroclásticos, etc).
5. Efectuar en otras zonas con calidad de datos mejor
6. Evaluar una ventana de frecuencias dependiendo el tipo de estudio

Bibliografía:

Cadena Ibarra, O. (2021). *Modelos de fuente de sismicidad LP para la actividad del volcán Galeras 2004-2010 (Colombia).* Universidad Nacional de Colombia.

Chong, Bao. (2021). K-means clustering algorithm: a brief review. Academic Journal of Computing & Information Science, 4(5), 37–40. https://doi.org/10.25236/AJCIS.2021.040506

Developments in Volcanology. (2003). Chapter 14 Seismic monitoring of volcanic activity and prediction of volcanic eruptions. Developments in Volcanology, 6(C), 235–252. https://doi.org/10.1016/S1871-644X(03)80214-0

Krischer, L., Megies, T., Barsch, R., Beyreuther, M., Lecocq, T., Caudron, C., & Wassermann, J. (2015). ObsPy: a bridge for seismology into the scientific Python ecosystem. Computational Science & Discovery, 8(1), 014003. https://doi.org/10.1088/1749-4699/8/1/014003

SGC. (2022). Boletín semanal de actividad del volcán Nevado del Ruiz del 08 al 14 de noviembre de 2022.

Watson, L. M. (2020). Using unsupervised machine learning to identify changes in eruptive behavior at Mount Etna, Italy. Journal of Volcanology and Geothermal Research, 405, 107042. https://doi.org/10.1016/J.JVOLGEORES.2020.107042

Zou, C. (2017). Meaning of Unconventional Petroleum Geology. Unconventional Petroleum Geology, 49–95. https://doi.org/10.1016/B978-0-12-812234-1.00002-9

Recursos web:

https://es.wikipedia.org/wiki/Similitud_coseno

https://es.wikipedia.org/wiki/Silhouette_(clustering)

https://www.elheraldo.co/economia/consejo-de-estado-ordena-suspender-produccion-de-gas-de-drummond-en-campo-la-loma-687184

https://medium.com/mlearning-ai/deciding-number-of-clusters-using-gap-statistics-davies-bouldin-index-calinski-harabasz-index-2ce9acfb6118#:~:text=Davies%20Bouldin%20index%20is%20calculated,cluster%20distance%20to%20cluster%20size.

https://towardsdatascience.com/cheat-sheet-to-implementing-7-methods-for-selecting-optimal-number-of-clusters-in-python-898241e1d6ad

http://sismo.sgc.gov.co:8080/fdsnws/dataselect/1/builder

https://stackoverflow.com/questions/18424228/cosine-similarity-between-2-number-lists