Caracterización de sismos en el norte de Chile utilizando técnicas de clustering para la identificación de posibles puntos de primera respuesta

Characterization of earthquakes in northern Chile using clustering techniques for the identification of possible first response points

Francisco García Barrera[1*] https://orcid.org/0000-0002-2411-9183
David Contreras Aguilar[1] https://orcid.org/0000-0002-6906-485X
Fernando Vergara Ramírez[1] https://orcid.org/0009-0006-9003-7019
Xenia Andaur Estica[1] https://orcid.org/0009-0000-1017-3552

Recibido 5 de abril de 2024, aceptado 03 de mayo de 2024
Received: April 05, 2024 Accepted: May 03, 2024

RESUMEN

En el presente artículo se explica el trabajo realizado para agrupar eventos sísmicos en el Norte Grande de Chile con el objetivo de identificar potenciales zonas geográficas donde se puedan desplegar posibles puntos de primera respuesta ante eventuales catástrofes que pudiesen ser provocadas por dichos eventos, lo anterior dada la alta actividad sísmica que tiene dicha zona. Para ello, se desarrolla el proceso de forma adaptada en base a las seis fases de la metodología CRISP-DM, dentro de la cual se aplican seis diferentes algoritmos de *clustering* para agrupar los datos, evaluando los resultados a través de diferentes índices de validación interna. En los resultados se obtiene un amplio espectro de escenarios dada la cantidad de algoritmos utilizados, los ajustes de los hiper parámetros aplicados, las cantidades de grupos configuradas y los distintos tipos de distancias empleadas. En particular, destacan los resultados obtenidos por medio de la técnica de *clustering* DBSCAN, la cual genera las mejores agrupaciones, identificando las localidades de Achuleo, norte de Pisagua, Pozo Almonte y Caleta Chipana como sectores candidatos a ser posibles puntos de primera respuesta. Finalmente se concluye que los algoritmos basados en densidad son los más recomendados para agrupar eventos sísmicos ya que tienen la capacidad de encontrar grupos con formas arbitrarias que no son necesariamente convexos.

Palabras clave: Minería de datos, sismos, *clustering*, DBSCAN, Norte Grande de Chile.

ABSTRACT

Based on the high seismic activity of Norte Grande de Chile, this article explains the work carried out to cluster seismic events to identify potential geographic areas where possible first response points can be deployed in case of catastrophes resulting from such events. The process is developed in an adapted form based on the six phases of the CRISP-DM methodology, within which six different clustering *algorithms are applied to group the data, evaluating the results through different internal validation indexes. The results show a wide range of scenarios given the number of algorithms used, the settings of the hyperparameters applied, the number of clusters configured, and the different types of distances used. Particularly noteworthy are the results obtained using the DBSCAN* clustering *technique, which generates the best groupings, identifying the localities of Achuleo, North of Pisagua, Pozo Almonte, and*

[1] Universidad Arturo Prat. Facultad de Ingeniería y Arquitectura. Iquique, Chile. Email: francgar@unap.cl;davicont@unap.cl; ferverga@estudiantesunap.cl; xandaur@unap.cl
* Autor de correspondencia: davicont@unap.cl

Caleta Chipana as candidate sectors to be possible first response points. It is concluded that density-based algorithms are the most recommended for clustering *seismic events since they can find clusters with arbitrary shapes that are not inherently convex.*

Keywords: Data mining, earthquakes, clustering*, DBSCAN, Norte Grande of Chile.*

INTRODUCCIÓN

La sismología corresponde al estudio de las ondas sísmicas, el cual proviene de casi 2000 años atrás y tiene su origen en China, donde su terminología se define como "seismos = sacudida" y "ology = estudio de" [1]. Chile es uno de los países a nivel mundial, más proclives a sufrir terremotos principalmente debido a que está ubicado justo al límite de la placa tectónica de Nazca, que limita con la placa Sudamericana [2]. Hoy en día son muchos los organismos y organizaciones nacionales que se preocupan de analizar y monitorear estos eventos, ya que cuando superan ciertas magnitudes podrían ocurrir catástrofes que deben ser atendidas de forma oportuna. Por lo anterior, es de interés conocer zonas geográficas que posean una baja cantidad de eventos sísmicos y con la menor magnitud posible para diferenciarlas de zonas que puedan tener una mayor concentración de estos eventos con altas magnitudes.

El norte de Chile históricamente no es ajeno a lo mencionado, ya que es muy recurrente que ocurran estos eventos en comparación con el resto del país [3]. Por otro lado, las técnicas de minería de datos (por ejemplo, el reconocimiento de patrones por métodos no supervisados) surgen como una alternativa que desempeña un papel dominante en las interpretaciones sismológicas basadas en datos sísmicos. Estas técnicas, se utilizan comúnmente para extraer (agrupar) información y patrones ocultos de una enorme cantidad de observaciones [4]-[8].

La propuesta de este trabajo consiste en recolectar y procesar los datos de los eventos sísmicos del norte de Chile, para posteriormente aplicar técnicas de minería de datos con el propósito de agrupar y caracterizar los sismos según las similitudes que estos puedan presentar. En concreto, el objetivo es identificar posibles Puntos de Primera Respuesta (PPR) encontrando zonas geográficas que agrupen sismos en bajas cantidades y con las menores magnitudes posibles para que instituciones puedan desplegar servicios de ayuda en dichas áreas, permitiendo una asistencia oportuna a zonas que pudiesen ser afectadas por una catástrofe de origen sísmico. Los modelos obtenidos con las técnicas de *clustering* son evaluados utilizando los índices de validación interna que permiten medir la separación y cohesión de los grupos formados por los métodos aplicados.

Las contribuciones que se derivan del presente artículo principalmente son: 1) Se presenta por primera vez un estudio de evaluación e identificación de puntos de primera respuesta aplicando técnicas de *clustering* para la evaluación de eventos sismológicos en el norte de Chile; 2) Se realizan experimentos basados en distintas técnicas de *clustering*, teniendo en cuenta análisis exploratorios, validación interna, estadísticas descriptivas y distribuciones de los grupos de sismos (clústeres) respecto a su magnitud y profundidad; y 3) Se identifican sectores geográficos donde hay una baja concentración de eventos sísmicos con magnitudes menores, los cuales podrían ser sectores candidatos para implementar posibles PPR en el norte de Chile frente a desastres causados por este tipo de fenómenos.

En la siguiente sección se exponen los principales trabajos relacionados. Posteriormente, se presenta la zona de estudio, metodología, técnicas de *clustering* e índices de validación. A continuación, se muestra configuración experimental y el análisis de resultados obtenidos. Finalmente, se presentan las conclusiones y trabajos futuros obtenidos del presente trabajo.

TRABAJOS RELACIONADOS

El trabajo [4], se enfoca en identificar los grupos geográficos que tienen eventos sísmicos similares al terremoto de Ecuador (Mw 7,8) considerando su latitud y longitud a partir del 16 de abril 2016 al 31 de mayo del 2018. Además, se identifican los eventos sísmicos centrales de cada grupo, asociándolo a una ubicación geográfica (ciudad más cercana) para

ayudar a los gobiernos en el desarrollo de políticas que permitan mejorar la mitigación de desastres después de un terremoto.

En este documento se presenta un análisis de conglomerados, a través del algoritmo MST-kNN utilizando distancia Haversin, donde los objetos agrupados son los eventos sísmicos (terremotos) considerando las coordenadas geográficas (latitudes y longitudes), profundidad y magnitud de tales eventos en la zona afectada. Como este algoritmo requiere una mínima intervención del usuario, se entrega al algoritmo la matriz de distancia previamente computada junto a los datos de los eventos sísmicos permitiendo realizar la agrupación de los 853 objetos en la zona afectada para encontrar grupos de terremotos con una ubicación geográfica similar. Una vez que se realiza el proceso de agrupación, se analizan los cúmulos de información contextual, es decir, los tipos de eventos sísmicos que contienen.

Una vez que el proceso de agrupación se realiza, se obtienen cinco agrupaciones geográficas, en donde se pueden observar que todos los eventos sísmicos frente a la costa del Ecuador están presentes en cada uno de los grupos. En conclusión, el grupo uno, dos y tres presentan eventos sismos concentrados en dos provincias del país en cuestión, permitiendo que la información generada por la investigación se pueda emplear como una herramienta para identificar el comportamiento de las zonas sísmicas que afectan áreas aledañas.

Al igual que el documento anteriormente nombrado, [5] escribe un artículo, donde habla de dos fallas geográficas activas en Indonesia. Aquí se presenta la agrupación del epicentro del terremoto en la provincia de Bengkulu. Los datos utilizados son desde los terremotos tectónicos en la provincia de Bengkulu y sus alrededores de enero de 1970 a diciembre de 2015.

En este análisis se utiliza el método de *K-Means* usando distancia Euclídea considerando la latitud, longitud y magnitud de los sismos. El número de clúster es determinado según el índice de validación Krzanowski and Lai (KL), el que después de un análisis exhaustivo los autores determinan que la cantidad de grupos que representa mejor la zona son 7 clústeres. En la variedad de grupos encontrados se mostró que cada uno lleva consigo características y un terremoto histórico, En el grupo número uno destaca que el 85,72% de los sismos son superficiales y el 14,28% restante son de mediana intensidad. El grupo dos se ubica la norte de Bengkulu siendo el terremoto del 2001 la magnitud más alta (7,4 Ms.). En el tercer clúster se agrupan 179 sismos, donde el 88.83% son entre 5 a 6 Ms. El grupo cinco se extiende al oeste de la provincia de Lampung, ocurren principalmente en el mar. En comparación con el anterior, el clúster cinco tiene una combinación de sismos generados en el mar y tierra. Finalmente, los grupos seis y siete son 100% terremotos ocurridos en el mar del Océano Indico.

Entre las conclusiones relevantes se descubrió que los terremotos en la provincia de Bengkulu y sus alrededores ocurren principalmente en el Océano Índico y solo los sismos pequeños ocurren en el continente. Como el mejor número de clúster según el índice KL es $K = 7$, el agrupamiento de terremotos con $K < 7$ trae consigo una gran variación en el agrupamiento, mientras que con $K > 7$ se producen agrupaciones superpuestas. Con el análisis de conglomerados *K-Means* se obtienen 5 grupos que se encuentran en el mar, la provincia continental de Bengkulu y alrededores de las Islas Mentawai.

Por otra parte, [6] muestra una clasificación de sismos con el método DBSCAN en el área de West Java, ubicada en las islas de Indonesia. Esta ubicación es muy inestable ya que limita con el Pacific Circum y Mediterranean Circum, esto atrae muchos sismos por volcanes activos y terremotos frecuentes. La población en este estudio son todos los eventos de terremotos que ocurrieron en el año 2021, esta investigación comenzó con el análisis del vecino más cercano para ver patrones de distribución de datos, cuando el patrón de distribución de datos ya está agrupado, se continuo con el análisis DBSCAN.

Los resultados del conglomerado se evalúan utilizando el coeficiente de silueta. Luego, en este estudio, se llevó a cabo una exploración de datos más profunda de tres maneras, a saber: (1) agrupamiento basado en el valor de silueta más alto, (2) agrupamiento al reducir el valor de MinPts y (3) agrupamiento basado en el límite superior más pequeño (supremo) del coeficiente de silueta. La exploración de datos aquí tuvo como objetivo formar la mayor cantidad de grupos sin dejar de considerar los límites del valor del coeficiente de

silueta para identificar más áreas propensas a los terremotos, pero también manteniendo la validez de los resultados obtenidos.

Los mejores resultados que se obtuvieron en la investigación fueron con Eps = 10000 y MinPts = 3, formando 12 grupos con un valor de coeficiente de silueta de 0,713, implicando que las agrupaciones tienen una estructura fuerte. Como conclusión los autores confirman que en la zona de West Java existen patrones de grupos sismos representados por los datos distribuidos, destacando el buen resultado que les entrego DBSCAN. Esperan que la información sobre la agrupación de áreas donde ocurren terremotos con frecuencia se pueda usar como una forma de mitigación de desastres sísmicos y minimizar el impacto de las pérdidas debido al terremoto.

La publicación [7]presenta un trabajo donde se agrupan de 1657 eventos sísmicos ocurridos en la India entre el 1 de enero de 2005 y el 31 de diciembre de 2015 considerando las variables de latitud, longitud, magnitud y profundidad. La agrupación de los datos se realiza con el objetivo de facilitar y apoyar la toma de decisiones usando k-means con la distancia euclidiana, los datos normalizados, y ajustando el número de clústeres e iteraciones, cuyo desempeño se evalúa con referencia a la suma de errores cuadrados dentro de los conglomerados.

Se obtiene como mejor resultado 6 grupos con 303, 149, 817, 218, 152 y 18 eventos sísmicos respectivamente. El trabajo revela que los grupos 2, 3 y 4 están más cerca mientras que los grupos 1, 3 y 6 están más lejos, concluyendo que k-means tiene el potencial de ser la herramienta adecuada para el análisis de clúster de terremotos.

En el estudio de microzonificación [8], se busca la clasificación de los peligros en una determinada zona en función de los movimientos superficiales del suelo que resultan de la amplificación y frecuencia de resonancia de estos frente a eventos sísmicos. Con el fin de reconocer los efectos locales sobre el suelo de la ciudad de Lahore en Pakistán, se obtuvieron los datos de amplitud (mm) y frecuencia (Hz) en 159 sitios implementando el método Nakamura de relación espectral horizontal a vertical (HVSR).

El algoritmo Fuzzy C-Mean (FCM) logró la mejor solución de agrupamiento utilizando los datos HVSR recolectados, siendo este evaluado con el Índice de silueta para demostrar que los modelos generados por FCM son los más consistentes. Los resultados revelan tres grupos importantes: Los clústeres 1 y 2 muestran que una parte importante del área de investigación posee niveles de frecuencias bajos a moderados (0,66 - 1,03 Hz) con una amplitud máxima de 2,25 - 4,38 mm, lo que indica la presencia de roca blanda a dura y espesor aluvial cubierta sedimentaria. El grupo 3 revela la presencia de rocas blandas a compactas (con frecuencias y amplitudes de 0,73 - 1,03 Hz y 3,02 - 4,11 mm, respectivamente) superpuestas al lecho rocoso. La ciudad de Lahore tiene el 60% de la cubierta del suelo con una amplitud de 2-3 mm (parte central) y alrededor del 40% de 3-4 mm al norte, sur y suroeste. Los mapas creados en este estudio proporcionan información útil relacionada con el movimiento del suelo esperado para reducir el riesgo sísmico para infraestructura en la ciudad de Lahore. Como conclusión, a pesar de los buenos resultados entregados por FCM, los autores recomiendan que es necesario un estudio adicional de variables, índice de vulnerabilidad, espesor de sedimentos, licuefacción de suelos, etc., en determinados lugares que revelan más de una característica del grupo.

En resumen, los artículos científicos analizados en su mayoría utilizan algoritmos de *clustering* particionales y tradicionales como K-means, además se concentran en la aplicación de solamente un algoritmo para encontrar un modelo descriptivo de su problemática. El presente trabajo busca aplicar una serie de algoritmos que se basan en diferentes estrategias para crear los grupos como lógica difusa, basados en densidad, particionales, entre otros, los cuales finalmente son evaluados con varios índices de validación entregando amplia gama de modelos y un proceso de evaluación robusto. Adicionalmente, la zona de estudio considerada para los experimentos agrega un particular contexto, ya que el norte de Chile se destaca por ser un área altamente sísmica despertando el interés por analizar estos eventos con el propósito de apoyar a la toma de decisiones ante eventuales desastres.

ZONA DE ESTUDIO

La zona geográfica que aborda la investigación considera una parte del Norte Grande de Chile, principalmente la región de Arica-Parinacota y Tarapacá. Las coordenadas de la zona se observan en

la Tabla 1, formando así un cuadrante que incluye las dos regiones antes mencionadas con algunas áreas territoriales del oeste de Bolivia y sur del Perú, lo cual se puede visualizar en la Figura 1.

Un estudio de vulnerabilidad sísmica de las ciudades del norte de Chile [9], aprecia que la historia sísmica de la zona de estudio está fuertemente influenciada por la ocurrencia de los grandes eventos de 1868 y 1879. Se establecieron algunas zonas de mayor potencial sísmico, destacando las regiones entre los paralelos 16 a 19 S y 20 a 22 S (conlleva gran parte del norte de Chile). Arica, se considera una región de madurez sísmica, que, por consiguiente, si se produjera un sismo de grandes proporciones, este mismo afectaría a Iquique y Tocopilla.

Tabla 1. Coordenadas geográficas zona de estudio.

Coordenada	Rango
Latitud	–17,4973894 a –21,47351753
Longitud	–68,34594727 a –70,3894043

Figura 1. Mapa zona de estudio.

Según [3], el norte de Chile y el sur del Perú son zonas conocidas mundialmente por la ocurrencia de grandes terremotos, registrando en el pasado grandes sismos y tsunamis con magnitudes cercanas a los 9 grados. Es por ello que, dado las zonas tectónicas y los últimos terremotos, la brecha sísmica se ve divida en estos dos sectores los cuales evidencian rupturas entre la placa de nazca y la sudamericana producto de grandes terremotos con poca profundidad y años de ocurrencia [2].

En Chile existen lagunas sísmicas o sitios donde no ha existido un sismo durante un largo periodo de tiempo (30 años), los cuales se vuelven difícil de identificar para tener una mejor estimación de la amenaza de estos fenómenos naturales producto de no poseer una red homogénea de estaciones sismológicas [10]. Cuando ocurren grandes terremotos estos liberan en su totalidad o gran parte la energía de una zona de ruptura, que, por consecuencia, se debe esperar un largo tiempo para que vuelva a ocurrir otro evento de similares magnitudes, ya que la zona debe recargar esa energía otra vez [10].

METODOLOGÍA

La investigación utiliza como marco de trabajo el ciclo de vida iterativo que establece la metodología CRISP-DM la cual considera en el nivel superior seis fases: Comprensión del Negocio, Comprensión de los Datos, Preparación de los Datos, Modelado, Evaluación y finalmente Despliegue. Cada una de estas fases tiene tareas genéricas, tareas especializadas e instancias de procesos estructuradas de forma jerárquica. En la Figura 2 se pueden ver las fases y la secuencia de ellas en el ciclo de vida, como también a continuación se describen en qué consiste cada una de ellas [11].

Comprensión del negocio

Esta fase inicial se centra en la comprensión de los objetivos y requisitos del proyecto desde una

Figura 2. Fases del modelo CRISP-DM [11].

perspectiva del dominio de la problemática, para luego convertir este conocimiento en una definición del problema de la minería de datos y un plan preliminar diseñado para lograr los objetivos.

Comprensión de los datos
La fase de comprensión de los datos comienza con la recopilación inicial de datos y continúa con actividades que permiten familiarizarse con los datos, identificar los problemas de calidad de los datos, descubrir las primeras percepciones de los datos y detectar subconjuntos interesantes para formar hipótesis sobre la información oculta.

Preparación de los datos
La fase de preparación de los datos abarca todas las actividades necesarias para construir el conjunto de datos final a partir de los datos brutos iniciales. Las tareas de esta etapa se realizan de forma iterativa y no necesariamente en el orden prescrito, las cuales incluyen la selección de tablas, registros y atributos, así como la transformación y limpieza de los datos para las herramientas de modelación.

Modelamiento
En la fase de modelado se seleccionan las técnicas y algoritmos más apropiados para el proyecto de minería de datos en función del problema a abordar, los datos disponibles, los requisitos del problema y el tiempo disponible con el objetivo de crear y refinar modelos apropiados al objetivo de minería de datos.

Evaluación
En esta etapa del proyecto, se busca evaluar los modelos generados de la etapa anterior en base a métricas de precisión, exactitud, error, compactación y separación (índices de validación), entre otros, según corresponda a los modelos empleados, con el propósito medir el cumplimiento de los criterios de éxito establecidos para los modelos y el problema. Siempre es pertinente revisar el proceso, considerando los resultados obtenidos para así, repetir y realizar refinamientos en algún paso anterior que permita mejorar el resultado de los modelos. Si el modelo generado es válido en función de los criterios de éxito establecidos en las fases anteriores y presenta métricas de evaluación aceptables para el problema, se procede a la explotación del modelo.

Despliegue
En la fase de despliegue, una vez que el modelo se ha validado y construido, el conocimiento se transforma en acciones dentro del proceso de negocio. Por lo general esta fase puede implicar acciones simples como la generación de un informe o complejas el desarrollo de aplicaciones para que los usuarios utilicen los modelos, la realización de mantenciones periódicas y automatizarlo dentro de una organización.

Algoritmos de Clustering
Para realizar la caracterización de los sismos, y así identificar posibles zonas de primera respuesta, los algoritmos de *clustering* se vuelven una buena opción para agrupar los sismos en base a la similitud que puedan presentan.

El agrupamiento o *clustering* es una de las tareas descriptivas empleadas por excelencia y consiste en obtener grupos "naturales" a partir de los datos. Estos datos son agrupados por las técnicas basándose en medidas similitud para crear grupos de datos lo más compacto posible y lo más distanciado de otros grupos [12]. Según los trabajos relacionados, las técnicas de *clustering* más utilizadas para agrupar sismos son: K-Means, K-Medoid, Fuzzy C-Means, Gaussian Mixture, Agrupamiento Jerárquico Aglomerativo y DBSCAN.

K-Means
Es un algoritmo de clasificación no supervisada de tipo particional, el cual está basado en la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuya distancia es menor. El número de grupos o clústeres, denotado como k, debe definirse antes de ejecutar el algoritmo, donde además cada grupo o clúster está definido por un punto, generalmente identificado como el centroide del clúster [13].

Este algoritmo funciona en dos fases principales: (1) Inicialización y (2) Asignación [13], donde la primera corresponde a la selección de los *k* centroides, lo cual puede ser de forma aleatoria o especifica de acuerdo con algún conocimiento previo de los datos. La segunda fase es iterativa y consiste en asignar a cada centroide los puntos de los conjuntos de datos más próximos formando así los *k* grupos de acuerdo con una medida de proximidad, la más utilizada es la distancia Euclidiana. Finalmente se recalculan los centroides utilizando la media en base a los puntos que forman parte de su grupo o partición. Estos dos pasos se repiten hasta que se satisface la condición de parada.

K-Medoid

Este algoritmo es una variante del algoritmo K-Means [13], el cual utiliza un dato en concreto como centroide (punto central) a diferencia de K-Means que emplea como centroide el promedio de los datos que componen un grupo. Esto permite que el algoritmo sea más robusto cuando las distribuciones son asimétricas o existen valores outliers. El algoritmo considera las dos fases de K-Mean, inicialización y asignación.

Fuzzy C-Means

Este algoritmo, Fuzzy C-Means (FCM), construye una partición difusa, donde inicialmente todos los puntos de datos tienen el mismo grado de pertinencia a los grupos, la cual es recalculada de forma iterativa en base a una función de pertenencia para finalmente asignar al dato el grupo con el cual tenga mayor peso [13].

Gaussian Mixture

Según [14], Gaussian Mixture (GM), es un algoritmo de agrupación que usa una función de densidad de probabilidad paramétrica representada como una suma ponderada de densidades de componentes gaussianos. Este modelo probabilístico es la combinación de múltiples distribuciones normales. Se puede entender como una generalización de *K-Means* donde en lugar de asignar cada observación a un único clúster, se obtienen distribuciones probabilísticas de pertenencia a cada uno.El proceso de *clustering* pasa a estimar los parámetros de la mezcla gaussiana mediante el algoritmo Expectation-Maximization (EM).

Agrupamiento jerárquico aglomerado

El algoritmo de agrupamiento aglomerativo jerárquico (AHC: Agglomerative Hierarchical Clustering) tiene dos enfoques, el primero considera inicialmente que todos los datos son un grupo individual, los cuales se van agrupando uno con otros para formar los clústeres, donde estos a la vez se unen con el grupo más cercano para formar un clúster más grande. El segundo corresponde al proceso inverso en el cual todos los datos son parte de un solo gran grupo el cual se va dividiendo en otros grupos más pequeños [13].

El objetivo es crear una secuencia de particiones anidadas, que se pueden visualizar a través de un árbol o una jerarquía de clústeres, también llamado dendrograma de clúster. El nivel más bajo del árbol (las hojas) consiste en cada punto en su propio grupo, mientras que el nivel más alto (la raíz) consiste en todos los puntos en un clúster (ambos se conocen como "agrupaciones triviales"). En algún nivel intermedio, se puede encontrar clústeres significativos [15].

El paso principal en el algoritmo es determinar el par más cercano de grupos, y para ello se pueden utilizar varios "criterios de proximidad", tales como "Single Link" (enlace único), "Complete Link" (enlace completo), "Average Link" (promedio de grupo), entre otros [15].

DBSCAN

"Density-Based Spatial Clustering of Applications with Noise" (DBSCAN), es un algoritmo especialmente diseñado para identificar grupos no convexos basado en densidad que produce un agrupamiento particional utilizando centroides, en el que el número de clústeres es determinado automáticamente por el algoritmo. Busca localizar regiones de alta densidad que están separadas entre sí por regiones de baja densidad, donde los puntos en regiones de baja densidad se clasifican como ruido y se omiten [16].

Según [15] el agrupamiento basado en densidad utiliza la densidad local de puntos para determinar los grupos, en lugar de utilizar sólo la distancia entre puntos. El algoritmo emplea un radio ε alrededor de un punto X llamado: ε – *vecindario* de X, donde X es un punto central ("Core Point") que puede tener una cantidad mínima de puntos (*minpts*) en su ε – *vecindario*, los cuales corresponden a un umbral de frecuencia o densidad definido por el usuario, al igual el ε. Un punto de borde ("Border Point") se

define como un punto que no cumple con el umbral de *minpts* pero pertenece al ε – *vecindario* de algún punto central *Z*. Por último, si un punto no es ni un núcleo ni un punto de frontera, entonces se llama un punto de ruido ("Noise Point") o un valor atípico.

Índices de validación para Clustering

Los modelos generados por los algoritmos de *clustering* normalmente son evaluados y validados utilizando índices de validación interna los cuales permiten medir que tan compactos y separados están los grupos formados. Entre más compactos y separados estén los grupos, el modelo es mejor [17]. Entre los índices de validación más utilizados se encuentran: Silueta, Davis-Bouldin, Dunn, Calinski-Harabasz y CDbw, los cuales se describen a continuación.

Índice de Silueta

El índice de silueta (SIL) [18] define la calidad de los resultados del *clustering* en función de la proximidad entre los objetos de un clúster concreto y la vecindad de estos objetos con el clúster más cercano, en términos más técnicos mide la cohesión y separación de los grupos y se basa en la diferencia de la distancia media de los puntos del clúster más cercano y los puntos de este. Un valor cercano a +1 indica que el dato está mucho más cerca de los puntos de su propio clúster y está lejos de otros clústeres. Un valor cercano a cero indica que el dato está cerca de la frontera entre dos clústeres. Por último, un valor cercano a –1 indica que, si está mucho más cerca de otro clúster que de su propio clúster, y, por tanto, el punto puede estar mal agrupado [15].

Índice de Davies-Bouldin

El índice Davies-Bouldin (DB) [19], es la relación entre la suma de la dispersión interna de los conglomerados y la distancia entre ellos. Cuanto menor sea el valor de DB, mejor será la agrupación, ya que significa que los clústeres están bien separados (es decir, la distancia entre las medias de los clústeres es grande), y cada clúster está bien representado por su media (es decir, tiene una pequeña dispersión) [15].

Índice de Dunn

El índice de Dunn [20], se calcula a partir de la relación entre la distancia intergrupal más corta y la distancia intragrupo más larga. Devuelve valores en el rango [0,∞), donde los valores más altos intentan identificar conglomerados compactos y bien separados [17].Cuanto mayor sea el índice de Dunn, mejor será la agrupación porque significa que incluso la distancia más cercana entre puntos de diferentes clústeres es mucho mayor que la distancia más lejana entre puntos del mismo clúster. Sin embargo, el índice de Dunn puede ser insensible porque las distancias mínimas inter clúster y máximas intracluster no capturan toda la información sobre una agrupación [15].

Índice de Calinski-Harabasz

El índice Calinski-Harabasz (CH) [21] es un índice popular que utiliza una relación ponderada de la suma de los cuadrados entre grupos (medida de separación) y la suma de cuadrados dentro del grupo (medida de compactación). Cuanto más alto el índice, mayor será el grado de separación entre los grupos formados y más compactos serán los datos dentro de ellos, es decir, mejor es el efecto de agrupación.

Índice de CDbw

CDbw (Composed Density between and within clusters) (17) tiene en cuenta todos los criterios de una "buena" agrupación (es decir, la cohesión de los clústeres, la compacidad y la separación), lo que permite una evaluación fiable de los resultados de la agrupación. Una agrupación con conglomerados compactos y bien separados y una baja variación de la distribución de la densidad dentro de los conglomerados da lugar a valores altos para los términos de CDbw (es decir, Cohesión, Separación y Compactación). Cuanto mayor sea el valor de CDbw corresponderá a un mejor agrupamiento de los datos.

CONFIGURACIÓN EXPERIMENTAL

El estudio utiliza los registros (Tabla 2) que proveen las principales plataformas sismológicas considerando la zona de estudio definida, estas son: Incorporated Research Institutions for Seismology (IRIS), United States Geological Survey (USGS) y el Centro Sismológico Nacional (CSN) de la Universidad de Chile (UChile).

Los datos extraídos de IRIS comprenden el período de enero del 2001 hasta agosto del 2020, USGS desde diciembre de 1918 hasta septiembre del 2020

y CSN desde enero del 2000 hasta agosto del 2020. En las tres fuentes de información se busca obtener los registros más antiguos disponibles hasta el año 2020, los cuales se encontraban en diferentes tipos de formatos, principalmente reportes y archivos Excel.

Cada fuente de información presenta una estructura diferente en los datos, IRIS presenta 13 atributos, CSN 7 atributos y USGS 22 atributos, por lo que en la fase de preparación de datos se procesan por separado para luego ser integradas como una única fuente de información considerando los atributos en común y más relevantes. Estos se encuentran indicados en la Tabla 3.

Una vez integrados los datos, estos son procesados para crear las vistas minables (VM) que alimentan a los algoritmos con el objetivo de crear los modelos. Para ello, se eliminan los eventos sísmicos duplicados que tengan la misma Latitud, Longitud, Fecha, Hora y Magnitud y Profundidad. Posteriormente, se descartan los eventos con magnitudes menores a 3 grados, profundidades menores a 0,5 Km y los que presenten valores vacíos. Una vez que los datos fueron procesados e integrados, el *dataset* presenta las estadísticas indicadas en la Tabla 4.

Tabla 2. Total de registros por cada fuente de información.

Fuentes de datos	Registros
IRIS	31607
CSN	109068
USGS	22641
Total	163316

Finalmente, se construyen cuatro VM las cuales consideran los siguientes atributos:

- **Vista Minable 1 (VM1):** Latitud, Longitud, Magnitud y Profundidad (datos estandarizados).
- **Vista Minable 2 (VM2):** Magnitud y Profundidad (datos estandarizados).
- **Vista Minable 3 (VM3):** Latitud, Longitud, Magnitud y Profundidad.
- **Vista Minable 4 (VM4):** Magnitud y Profundidad.

Los atributos fecha, hora, tipo de magnitud y ubicación, no fueron utilizados en las VM, ya que no aportan información relevante para los algoritmos de *clustering* que serán utilizados.

La estandarización de los datos en VM1 y VM 2 se realiza con el método Min-Max, cuya fórmula corresponde a la ecuación (1) [22] permitiendo trasformar los datos a un rango de 0 hasta el 1. Donde X_i corresponde al dato que se estandariza, *min (X)* corresponde al menor valor de la serie de datos, y *max (X)* corresponde al máximo valor de la serie.

$$X_{new} = \frac{X_i - \min(X)}{\max(X) - \min(X)} \quad (1)$$

Tabla 3. Atributos seleccionados e integrados.

Atributo	Descripción
Fecha	Fecha del evento.
Hora	Hora del evento.
Latitud	Latitud del evento en número.
Longitud	Longitud del evento en número.
Profundidad	Profundidad del evento en kilómetros.
Magnitud	Magnitud del evento.
Tipo de magnitud	Escala del evento.
Ubicación	Ubicación geográfica del evento.

Tabla 4. Estadísticos descriptivos del *Dataset*.

Variable	mean	std	min	Q1	Mediana	Q3	max
Magnitud	3,6948	0,6509	3	3,2	3,5	4,1	7,9
Profundidad (Km)	104,5753	28,8264	0,6	95,9	108,1	120,6	269,1

En los experimentos, las cuatro VM son consumidas por los algoritmos *K-Means*, *K-Medoid*, FCM, GM, AHC y DBSCAN los cuales a la vez son configurados con tres métricas de distancias diferentes: Coseno, Euclidiana y Manhattan. Estas últimas son aplicadas con todos los algoritmos exceptuando la distancia de Coseno con DBSCAN dado que se recomienda utilizar Euclidiana y Manhattan [23].

La cantidad experimentos y modelos generados está dada por la combinatoria de aplicar los seis algoritmos mencionados utilizando las cuatro VM con las tres métricas de distancia y la cantidad de grupos que fueron configurados (de 2 a 20 grupos). Esto entrega un total de 228 modelos por algoritmo, exceptuando DBSCAN el cual entrega la cantidad de grupos de forma natural dado los hiper parámetros de ε y *minpts*. La Tabla 5 muestra el detalle de los hiper parámetros utilizados para cada uno de los algoritmos.

Los modelos son evaluados con los índices de validación interna: SIL, DB, Dunn, CH y CDbw, permitiendo observar el grado de compactación y separación de los grupos generados. El índice de Dunn no es aplicado con DBSCAN dado los resultados en la comparativa de índices de validación de [24] para los algoritmos basados en densidad.

También se despliegan en mapas para visualizar geográficamente los clústeres de los modelos clasificando los sismos según la Tabla 6 para facilitar la identificación de los posibles puntos de primera respuesta. En la Figura 3 se presenta un mapa con la distribución geográfica de los sismos que son utilizados para crear las VM y los grupos por medio de los algoritmos.

Todo el proceso se desarrolla con la herramienta KNIME y Python las cuales son integradas y permiten construir flujos de trabajos semiautomatizados para abordar desde la lectura de los datos, la integración, limpieza/transformación, creación de VM, hasta la implementación y aplicación de los algoritmos

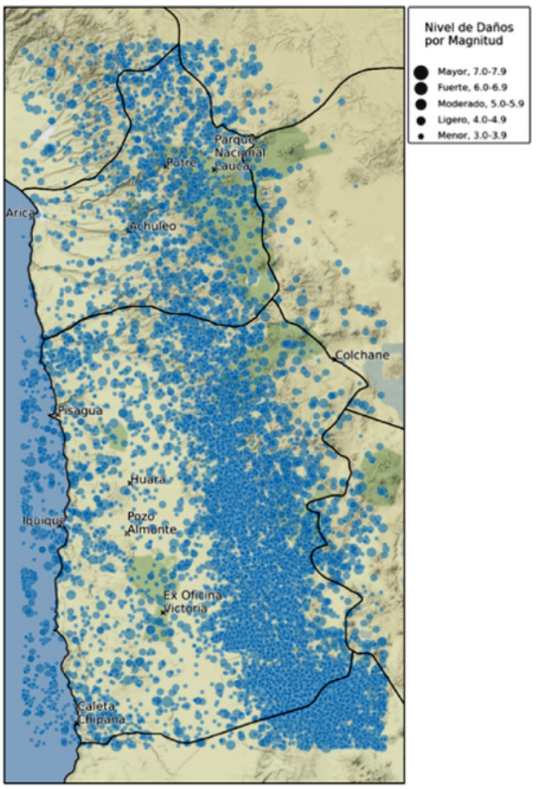


Figura 3. Mapa de eventos sísmicos en la zona de estudio.

Tabla 6. Clasificación sismos según magnitud.

Clasificación	Magnitud
Cataclismo	8 o mas
Mayor	7 - 7.9
Fuerte	6 - 6.9
Moderado	5 - 5.9
Ligero	4 - 4.9
Menor	3 - 3.9

Tabla 5. Hiper parámetros utilizados para los experimentos.

Algoritmo	Iteraciones	K	ε	*minpts*	Distancia
KMeans	300	2 a 20	N/A	N/A	Coseno, Euclidiana, Manhattan
KMedoid	300	2 a 20	N/A	N/A	Coseno, Euclidiana, Manhattan
FCM	100	2 a 20	N/A	N/A	Coseno, Euclidiana, Manhattan
AHC	100	2 a 20	N/A	N/A	Coseno, Euclidiana, Manhattan
DBSCAN	N/A	N/A	0,05 - 0,08	10, 16, 20, 32	Euclideana y Manhattan
GM	100	2 a 20	N/A	N/A	Coseno, Euclidiana, Manhattan

de minería de datos para obtener los modelos y la evaluación de estos.

ANÁLISIS DE RESULTADOS

En esta sección se visualizan los resultados obtenidos utilizando los algoritmos exponiendo los índices de validación interna que permiten revelar información importante de los distintos modelos generados con sus respectivos mapas. Para lograr un mayor espectro de resultados se realizaron experimento desde 2 a 20 clústeres por cada VM y métricas de distancias (explicado en el punto anterior). Adicionalmente, cabe destacar que en todos los experimentos se realizaron análisis con estadísticos descriptivos y visualizaciones de distribución para caracterizar los grupos formados en los modelos, sin embargo, dado los extensos resultados por el amplio espectro de algoritmos, VM, ajustes de hiper parámetros y métricas de distancias empleadas, dichas estadísticas y visualizaciones se muestran solamente para el algoritmo DBSCAN ya que entregó los mejores resultados.

Resultados *K-Means*

A continuación, se presentan los índices de validación interna y mapa generado del mejor modelo del algoritmo mencionado.

Índices de validación interna y mapa generado

En la Tabla 7 se muestran los valores de los índices de validación interna con las diferentes cantidades clústeres configuradas (2 a 20 grupos) para la VM1 usando la distancia euclidiana ya que entregaron los mejores resultados entre las distintas VM, matrices de distancias y ajustes de hiper parámetros con el algoritmo *K-Mean*. Se puede apreciar que el modelo de 3 grupos presenta los resultados más favorables en base a los índices de SIL, DB, Dunn y CH. Por otra parte, en los modelos mayores a 3 grupos los índices de validación tienden a ser menos favorables indicando que la calidad de los clústeres va disminuyendo mientras aumenta el número de grupos.

En la Figura 4 se aprecia un mayor detalle de los eventos agrupados con sus niveles de daños por magnitudes. Específicamente, se identifica de manera muy clara los 3 grupos bien formados concentrándose la mayor cantidad de eventos en el clúster 0 (6931 datos). Sin desmedro de lo anterior, no es posible

Tabla 7. Índice de validación interna para *K-Means* con VM1 y distancia Euclidiana.

Grupos	SIL	DB	Dunn	CH
2	0,4301	1,0234	0,3505	8893,361
3	0,4584	0,9466	0,4125	8091,238
4	0,3527	1,0242	0,2545	7998,944
5	0,2810	1,1118	0,1776	7102,158
6	0,2911	1,0843	0,1751	6650,153
7	0,2937	1,1052	0,1763	6447,209
8	0,2900	1,1017	0,1761	6130,544
9	0,2763	1,1383	0,1663	5868,554
10	0,2733	1,1400	0,1607	5639,120
11	0,2709	1,1516	0,1757	5401,464
12	0,2720	1,1330	0,1839	5202,343
13	0,2770	1,1480	0,2192	5055,013
14	0,2554	1,1820	0,1652	4913,381
15	0,2570	1,1577	0,1302	4800,189
16	0,2568	1,1819	0,1346	4637,926
17	0,2507	1,1849	0,1569	4509,887
18	0,2548	1,1874	0,1581	4406,914
19	0,2533	1,1816	0,1569	4322,177
20	0,2556	1,1673	0,1570	4223,984

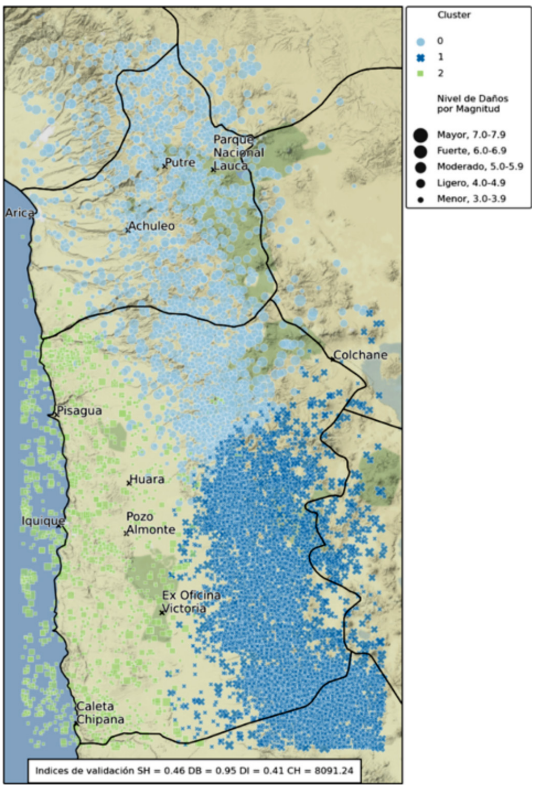


Figura 4. Modelo de 3 grupos con *K-Means* distancia Euclidiana y VM1.

identificar claramente posibles PPRs, ya que los clústeres formados por el algoritmo son muy grandes y demasiado densos para encontrar zonas geográficas acotadas y focalizadas que tengan una baja ocurrencia de sismos y con las menores magnitudes posibles.

Resultados K-medoid

Como se pudo observar en el algoritmo *K-Means*, el mejor modelo corresponde a la formación de 3 clústeres. En ese caso, se observa que se logra agrupar hasta 2 grandes zonas del Norte Grande ya que al aumentar la cantidad clústeres, estos bajan la calidad de los índices de validación. A continuación, se presentan los mejores resultados obtenidos utilizando el algoritmo *K-Medoid.*

Índices de validación interna y mapa generado

En la Tabla 8 se muestran los valores de los índices de validación interna para la VM1 usando la distancia euclidiana ya que entregaron los mejores resultados entre las distintas VM, matrices de distancias y ajustes de hiper parámetros con el algoritmo *K-Medoid*. Se puede apreciar que el modelo de 2 grupos presenta los resultados más favorables dado que valores superiores reducen la calidad de las agrupaciones formadas.

Tabla 8. Índice de validación interna para *K-Medoid* con VM1 y distancia Euclidiana.

Grupos	SIL	DB	Dunn	CH
2	0,4232	1,0195	0,3490	8813,008
3	0,2867	1,2548	0,2107	6754,137
4	0,3450	1,0270	0,2476	7972,625
5	0,3263	1,1163	0,2602	6785,445
6	0,2753	1,1663	0,1874	6443,519
7	0,2903	1,1210	0,2005	6302,782
8	0,2667	1,2078	0,2133	5762,138
9	0,2718	1,2399	0,1998	5687,194
10	0,2698	1,2919	0,2013	5292,656
11	0,2605	1,2221	0,2084	5203,219
12	0,2374	1,2267	0,1520	4982,505
13	0,2080	1,2802	0,1134	4664,119
14	0,2002	1,2968	0,1145	4465,287
15	0,2075	1,2478	0,1144	4458,319
16	0,2140	1,2139	0,1217	4349,350
17	0,2081	1,2432	0,1213	4200,643
18	0,2041	1,2568	0,1046	4107,139
19	0,2044	1,2367	0,1046	4056,801
20	0,2049	1,2120	0,1035	3995,960

Adicionalmente, se puede observar en el mapa de la Figura 5 mayor detalle los eventos agrupados con sus niveles de daños por magnitudes. Específicamente, se identifica de manera muy clara los 2 grupos formados concentrándose la mayor cantidad de eventos en el clúster 1 (6961 datos), no logrando así identificar los posibles PPR ya que los clústeres formados por el algoritmo son muy grandes y demasiado densos para encontrar zonas geográficas acotadas y focalizadas que tengan baja ocurrencia de sismos y con las menores magnitudes posibles.

Resultados Fuzzy C-Means

En esta sección se presentan los índices de validación interna y mapa generado de los mejores resultados obtenidos utilizando el algoritmo *FCM*.

Índices de validación interna y mapa generado

En la Tabla 9 se puede apreciar que, de acuerdo con los índices de validación interna, los resultados más favorables se obtienen con el modelo de 2 clústeres. Lo anterior se logra utilizando la VM1 y la distancia

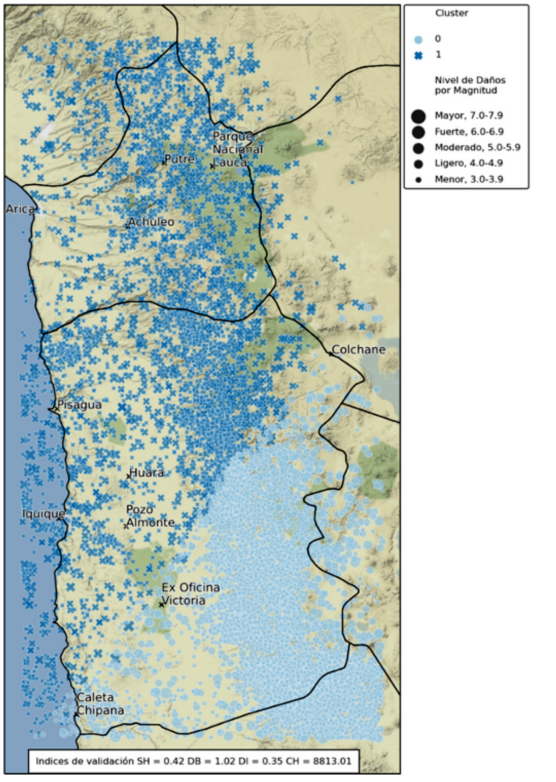


Figura 5. Modelo de 2 grupos con *K-Medoid* distancia Euclidiana y VM1.

Tabla 9. Índice de validación interna para FCM con VM1 y distancia Euclidiana.

Grupos	SIL	DB	Dunn	CH
2	0,4254	1,0224	0,3487	8870,787
3	0,3191	1,2719	0,2430	7083,327
4	0,3362	1,0376	0,2378	7924,070
5	0,2698	1,1591	0,1715	7020,901
6	0,2793	1,1381	0,1772	6530,597
7	0,2587	1,2021	0,1665	6033,542
8	0,2674	1,1437	0,1647	6021,675
9	0,2277	1,2388	0,1161	5568,535
10	0,2390	1,2113	0,1182	5425,733
11	0,2393	1,2113	0,1159	5243,432
12	0,2383	1,2864	0,1143	4994,523
13	0,2340	1,3265	0,1105	4743,375
14	0,2283	1,2837	0,1165	4558,726
15	0,2276	1,3381	0,1062	4463,651
16	0,2205	1,3887	0,1027	4246,083
17	0,2240	1,2848	0,1170	4177,023
18	0,2016	1,3841	0,0972	3986,418
19	0,1967	1,3819	0,0966	3856,243
20	0,2039	1,3196	0,1014	3885,516

euclidiana ya que entregan los mejores resultados entre todas las VM y matrices de distancias. Modelos mayores a 2 grupos provocan que las agrupaciones tiendan a bajar su calidad, lo cual se puede visualizar en los índices.

Adicionalmente, se puede observar en el mapa de la Figura 6 un mayor detalle los eventos agrupados con sus niveles de daños por magnitudes. Específicamente, se identifica de manera muy clara los 2 grupos formados concentrándose la mayor cantidad de eventos en el clúster 1 (6912 datos), no logrando así identificar los posibles PPR ya que los clústeres formados por el algoritmo son muy grandes y demasiado densos para encontrar zonas geográficas acotadas y focalizadas que tengan baja ocurrencia de sismos y con las menores magnitudes posibles.

Resultados agrupamiento jerárquico aglomerado
En esta sección se presentan los índices de validación interna y mapa generado de los mejores resultados obtenidos utilizando el algoritmo *AHC*.

Índices de validación interna y mapa generado
En la Tabla 10 se puede apreciar que, de acuerdo con los índices de validación interna, los resultados más

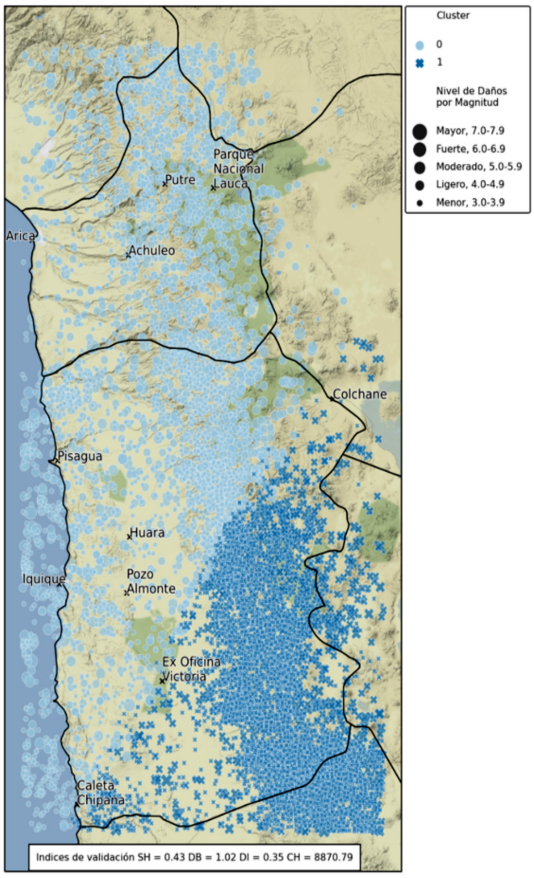


Figura 6. Modelo de 2 grupos con FCM distancia Euclidiana y VM1.

favorables se obtienen con el modelo de 3 grupos. Lo anterior se logra utilizando la VM1 y la distancia euclidiana ya que entregan los mejores resultados entre todas las VM y matrices de distancias. Modelos mayores a 3 grupos provocan que las agrupaciones tiendan a bajar su calidad, lo cual se puede visualizar en los índices.

Adicionalmente, se puede observar en el mapa de la Figura 7 un mayor detalle los eventos agrupados con sus niveles de daños por magnitudes. Específicamente, se identifica de manera muy clara los 3 grupos formados concentrándose la mayor cantidad de eventos en el clúster 1 (5882 datos), no logrando así identificar los posibles PPR ya que los clústeres formados por el algoritmo son muy grandes y demasiado densos para encontrar zonas geográficas acotadas y focalizadas que tengan baja ocurrencia de sismos y con las menores magnitudes posibles.

Tabla 10. Índice de validación interna para AHC con VM1 y distancia Euclidiana.

Grupos	SIL	DB	Dunn	CH
2	0,3880	1,0365	0,3318	8356,891
3	0,4022	1,1033	0,3666	7269,163
4	0,3531	1,0555	0,3042	7117,722
5	0,2913	1,2280	0,1873	6368,788
6	0,2360	1,1488	0,1608	5815,411
7	0,2306	1,1921	0,1608	5489,253
8	0,2329	1,2715	0,1608	5243,688
9	0,2266	1,3018	0,1707	4950,976
10	0,2296	1,2916	0,1707	4752,762
11	0,2301	1,3559	0,2049	4479,614
12	0,2288	1,3278	0,2067	4267,614
13	0,2251	1,3616	0,2099	4095,033
14	0,1874	1,3881	0,1493	3959,658
15	0,1870	1,4097	0,1493	3842,199
16	0,1634	1,4017	0,1493	3745,428
17	0,1726	1,4188	0,1311	3651,941
18	0,1736	1,3468	0,1311	3577,855
19	0,1757	1,3288	0,1311	3497,982
20	0,1783	1,3616	0,1311	3424,576

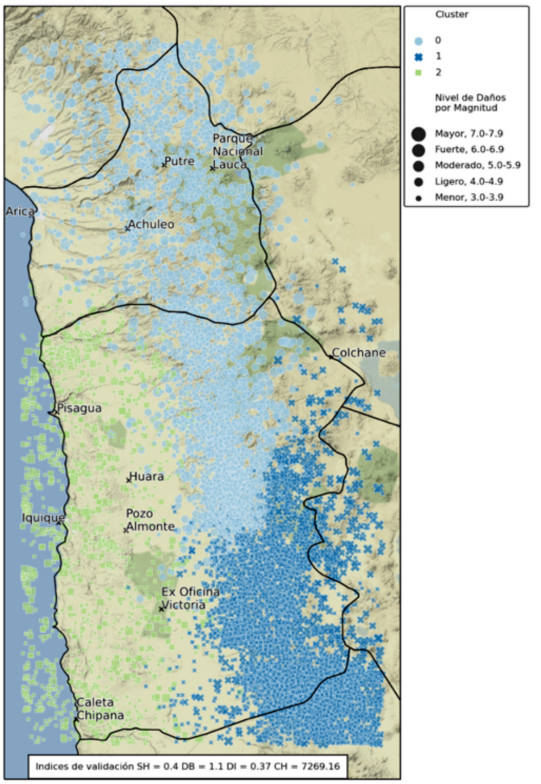


Figura 7. Modelo de 3 grupos con AHC distancia Euclidiana y VM1

Resultados Gaussian Mixture

En esta sección se presentan los índices de validación interna y mapa generado de los mejores resultados obtenidos utilizando el algoritmo *GM*.

Índices de validación interna y mapa generado

En la Tabla 11 se puede apreciar que, de acuerdo con los índices de validación interna, los resultados más favorables se obtienen con el modelo de 3 grupos. Lo anterior se logra utilizando la VM1 y la distancia euclidiana ya que entregan los mejores resultados entre todas las VM y matrices de distancias. Modelos mayores a 3 grupos provocan que las agrupaciones tiendan a bajar su calidad, lo cual se puede visualizar en los índices.

Adicionalmente, se puede observar en el mapa de la Figura 8 un mayor detalle los eventos agrupados con sus niveles de daños por magnitudes. Específicamente, se identifica de manera muy clara los 3 grupos formados concentrándose la mayor cantidad de eventos en el clúster 2 (6150 datos), no logrando así identificar los posibles PPR ya que los clústeres formados por el algoritmo son muy grandes y demasiado densos para encontrar zonas geográficas acotadas y focalizadas que tengan baja ocurrencia de sismos y con las menores magnitudes posibles.

Tabla 11. Índice de validación interna para GM con VM1 y distancia Euclidiana.

Grupos	SIL	DB	Dunn	CH
2	0,3721	1,2496	0,3062	6037,319
3	0,3646	1,2035	0,3002	4948,893
4	0,2296	1,4368	0,1910	4215,513
5	0,2185	1,6092	0,1790	3708,703
6	0,1837	1,5114	0,1692	3749,996
7	0,1909	1,3780	0,1705	4372,671
8	0,1665	2,1368	0,1042	3489,614
9	0,1642	2,1523	0,0962	3418,599
10	0,1396	2,4866	0,069	3151,417
11	0,1137	2,5733	0,0673	3043,971
12	0,1142	2,6604	0,0591	3056,811
13	0,0891	2,2521	0,0737	2712,083
14	0,0992	2,2686	0,0703	2698,089
15	0,1115	2,0655	0,0743	2641,454
16	0,1133	1,8123	0,0889	2763,205
17	0,1013	1,9368	0,0804	2504,296
18	0,0887	1,9816	0,0715	2384,655
19	0,1046	1,7311	0,0621	2509,956
20	0,1019	1,8465	0,0721	2401,164

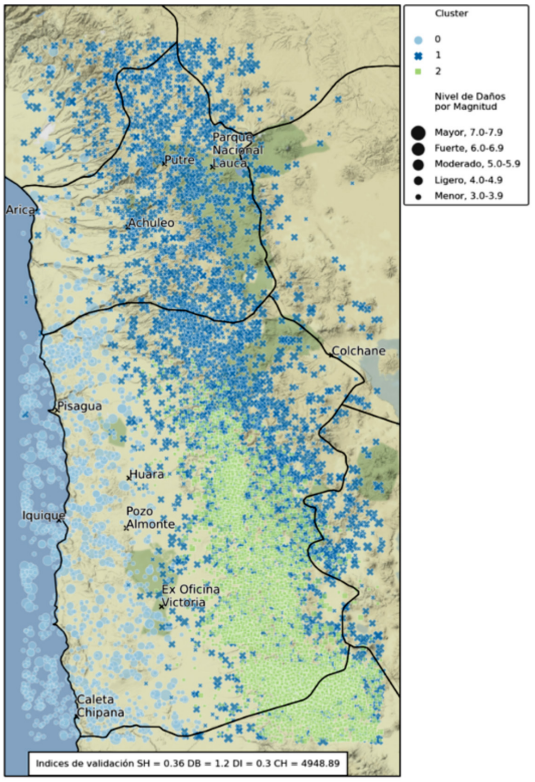


Figura 8. Modelo de 3 grupos con GM distancia Euclidiana y VM1.

Resultados DBSCAN

En esta sección se presentan los índices de validación interna, estadísticas descriptivas, gráficos y mapa generado de los mejores resultados obtenidos utilizando el algoritmo *DBSCAN*.

Índices de validación interna y mapa generado

En la Tabla 12 se pueden observar los índices de validación interna, número de grupos y cantidad de datos considerados como ruido por el algoritmo para los ε que presentaron los resultados más favorables con *minpts* = 16 utilizando la VM1 y la distancia euclidiana, ya que estos últimos elementos mostraron ser los mejores ajustes de hiper parámetros, VM y matriz de distancia en los experimentos. Se destaca el modelo con ε = 0,058ε el cual logra el mejor resultado formado 12 grupos de forma natural.

A diferencia de los modelos obtenidos con los algoritmos anteriores, los cuales crean agrupaciones muy grandes y demasiado densas que impedían identificar zonas geográficas acotadas y focalizadas para posibles PPR, al observar el mapa de la Figura 9, se visualiza que DBSCAN puede crear grupos más diferenciados con bajas cantidades de sismos y

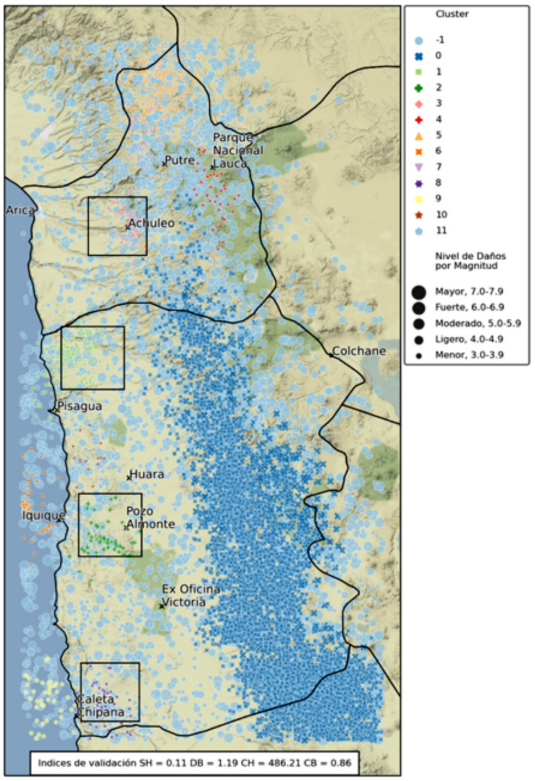


Figura 9. Modelo de 12 grupos con DBSCAN (Eps = 0,058) (*MinPts* =16) distancia Euclidiana y VM1.

Tabla 12. Índice de validación interna para DBSCAN con VM1 y distancia Euclidiana.

Épsilon (ε)	SIL	DB	CH	CDbw	Nº Grupos	Nº con Ruidos
0,058	0,1149	1,1896	486,2141	0,8574	12	3374
0,059	0,1197	1,1967	498,1144	0,7976	12	3288
0,060	0,1278	1,2090	604,7237	0,6129	10	3208
0,061	0,0916	1,3341	650,4706	0,5078	9	3083
0,062	0,0294	1,3843	532,3222	0,6496	8	2932

de pequeñas magnitudes, lo cual es atribuido a su capacidad de encontrar grupos con formas arbitrarias que no son necesariamente convexos.

Dado lo anterior, se identifican cuatro posibles PPR destacados con un cuadrado que encierra el área. Estos puntos corresponden a las zonas de: Achuleo (Grupo 3), norte de Pisagua (Grupo 1), Pozo Almonte (Grupo 2) y Caleta Chipana (Grupo 8).

Específicamente, en estas localidades del norte de Chile sería recomendable que exista una red de seguridad ante un eventual sismo de mayor grado y genere algún peligro mayor a sus alrededores. También se puede observar cómo los sismos que se encuentran dentro de los PPR identificados en su mayoría son bastantes leves dados el nivel de daños por magnitud bajo el cual fueron clasificados. A continuación, se presentará las estadísticas descriptivas y gráficos para ir más a detalle en estos PPR.

Estadísticas descriptivas y distribuciones

La Tabla 13 muestra las estadísticas descriptivas para las magnitudes y profundidades de cada clúster, como también la cantidad de eventos sísmicos que tiene cada uno (columna count), adicionalmente la Figura 10 muestra gráficamente las distribuciones de las variables mencionadas. En concreto, de las figuras y gráficos presentados es posible observar:

a. Teniendo en cuenta los cuatro posibles PPR, estos muestran tener bajas concentraciones de sismos con pequeños niveles de magnitudes promedio y valores máximo también bajos. Otros clústeres tienen características similares, sin embargo, la ubicación geográfica de los cuatro PPR es mejor ya que no se encuentran tan cercanos a la cordillera y tampoco a la costa. La localidad de Pozo Almonte que representa el clúster 2 tiene eventos bastante bajos identificando una zona bien prometedora para generar una red de seguridad ante un eventual peligro sísmico para la zona.
b. Con respecto a las *Profundidades* de los eventos sísmicos, estas son relativamente bajas en los cuatro PPR, destacando el caso de la localidad de Pozo Almonte el cual además tiene poca concentración de dichos eventos, permitiendo que al ser más superficiales y de baja magnitud no sean tan devastadores, por lo que es bastante ventajoso una zona con dichas características.
c. La cantidad de sismos para cada clúster identificado como posible PPR (*) son bajas por lo que resultan ventajosas zonas de estas características como la localidad de Pozo Almonte, siendo esta última la más recomendado para un posible PPR (Tabla 13).
d. En el 50% de los eventos sísmicos, las *magnitudes* del clúster 2 (localidad de Pozo Almonte) se ubican de 3,1 a 3,4 grados, donde su máximo valor tiene un registro de 3,7 grados siendo una zona candidata para ser un posible PPR.
e. En el 50% de los eventos sísmicos, las *profundidades* del clúster 2 (localidad de Pozo Almonte) se ubican de 48 a 54 kilómetros, volviéndolo una zona geográfica interesante para desplegar servicios que apoyen las áreas aledañas.
f. El clúster 1, 3 y 8 tienen características similares (Tabla 13), ya que presentan bajas concentraciones de sismos con pocas profundidades y magnitud. Características que vuelven dichas zonas geográficas interesantes para instalar posibles puntos de primera respuesta.

CONCLUSIONES

En el presente trabajo se ha logrado identificar zonas geográficas en el Norte Grande de Chile que agrupan baja cantidad de eventos sísmicos con magnitudes pequeñas, las cuales podrían ser zonas candidatas para instalar posibles PPRs permitiendo una asistencia oportuna a zonas que pudiesen ser afectadas por una catástrofe de origen sísmico.

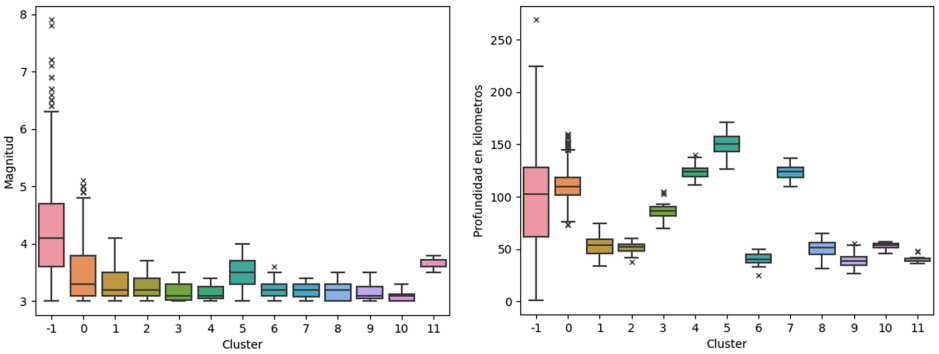
En concreto, se ha desarrollado el proceso de Knowledge Discovery in Databases (KDD) implementado la metodología CRISP-DM para entender el problema y desarrollar la solución propuesta. Específicamente, se identificaron y analizaron las diferentes fuentes de información sísmicas recolectadas, como IRIS, CSN y el USGS con registros de distintos periodos temporales de acuerdo con la zona geográfica a analizar.

Gracias a la plataforma KNIME, se logró un proceso semiautomático para la limpieza y transformación de los datos sismológicos para las tres fuentes de datos, desde la obtención de estos mismos, la integración, limpieza/transformación, creación de VM hasta la implementación y aplicación de los

Tabla 13. Estadísticas descriptivas para los 12 clústeres formados con DBSCAN (*Eps* = 0,058) (*MinPts* = 16) distancia euclidiana y VM1.

Magnitud										
Clúster	**count**	**mean**	**std**	**min**	**Q1**	**median**	**Q3**	**max**	**skew**	**kurtosis**
–1	3374	4,1663	0,7562	3,0	3,6	4,1	4,7	7,9	0,4569	0,0132
0	6970	3,5006	0,4724	3,0	3,1	3,3	3,8	5,1	0,9327	–0,1079
1*	137	3,3343	0,3050	3,0	3,1	3,2	3,5	4,1	0,8523	–0,4151
2*	59	3,2492	0,2046	3,0	3,1	3,2	3,4	3,7	0,4144	–0,8808
3*	34	3,1618	0,1498	3,0	3,0	3,1	3,3	3,5	0,7307	–0,4476
4	39	3,1641	0,1386	3,0	3,1	3,1	3,3	3,4	0,4247	–0,9948
5	121	3,5041	0,2650	3,0	3,3	3,5	3,7	4,0	–0,0609	–0,8957
6	43	3,2419	0,1622	3,0	3,1	3,2	3,3	3,6	0,3509	–0,7006
7	36	3,1722	0,1406	3,0	3,1	3,2	3,3	3,4	0,3173	–1,0949
8*	53	3,1547	0,1514	3,0	3,0	3,2	3,3	3,5	0,4298	–1,0722
9	43	3,1674	0,1507	3,0	3,1	3,1	3,3	3,5	0,7364	–0,3604
10	16	3,0875	0,0957	3,0	3,0	3,1	3,1	3,3	0,7208	–0,5203
11	12	3,6500	0,1087	3,5	3,6	3,6	3,7	3,8	0,2217	–1,1716
Profundidad (Km)										
Clúster	**count**	**mean**	**std**	**min**	**Q1**	**median**	**Q3**	**max**	**skew**	**kurtosis**
–1	3374	95,939	41,834	0,6	61,5	102,07	127,6	269,1	–0,0977	–0,6688
0	6970	110,754	13,145	73,0	101,4	109,25	118,7	160,0	0,4844	0,0894
1*	137	53,305	8,822	34,0	46,0	53,70	59,3	74,2	–0,0144	–0,5746
2*	59	51,227	5,043	37,9	48,0	52,10	54,5	60,2	–0,3874	–0,2125
3*	34	86,232	7,407	70,0	81,7	86,05	90,0	104,9	0,7940	1,0352
4	39	123,610	6,523	111,5	119,3	123,60	127,2	139,8	0,3118	–0,0307
5	121	150,154	9,374	126,4	143,2	150,00	157,4	171,1	–0,1156	–0,6756
6	43	40,761	5,195	24,9	37,3	40,40	45,3	49,9	–0,4263	0,4079
7	36	123,013	7,151	109,5	118,6	123,85	128,3	137,0	–0,3019	–0,6859
8*	53	49,793	7,685	31,4	45,0	51,10	55,7	64,7	–0,4012	–0,4118
9	43	38,884	6,280	26,1	34,9	38,30	42,7	55,0	0,5286	0,4861
10	16	52,656	3,195	45,6	50,9	53,70	55,0	56,6	–0,7479	–0,4009
11	12	40,742	3,813	35,8	38,5	40,65	41,3	48,4	0,8531	–0,0447

(*) Representan el clúster identificado como posible PPR del mapa visualizado en la Figura 9.
(–1) Corresponde a los datos considerados como ruido (Noise) por el algoritmo, es decir, no son un grupo.



(a) (b)

Figura 10. a) Distribución de magnitud y b) Profundidad para 12 clústeres formados con DBSCAN (*Eps* = 0,058) (*MinPts* = 16) usando distancia euclidiana y VM1.

algoritmos de minería de datos para obtener los modelos y la evaluación de estos.

Se utilizaron distintas técnicas de *clustering* que permitieron agrupar los eventos sísmicos según sus similitudes para identificar zonas geográficas. Lo anterior se logró gracias a la evaluación de varios algoritmos que utilizan diferentes estrategias para crear grupos, los ajustes de los hiper parámetros aplicados, las cantidades de grupos configurados y los distintos tipos de distancias empleadas.

A través del análisis de los resultados obtenidos en los experimentos, se logró obtener un amplio espectro de escenarios con diferentes recomendaciones de acuerdo con sus índices de validación interna. El algoritmo DBSCAN obtuvo los mejores resultados identificando las localidades de Achuleo, norte de Pisagua, Pozo Almonte y Caleta Chipana como sectores candidatos a ser puntos de primera respuesta (localidades que incluyen tanto sectores costeros como del interior del norte de Chile). Finalmente, se concluye que los algoritmos basados en densidad como DBSCAN son los más recomendados para agrupar eventos sísmicos ya que tienen la capacidad de encontrar grupos con formas arbitrarias que no son necesariamente convexos.

A pesar de que en los experimentos se utilizaron cuatro VM diferentes y tres métricas de distancias distintas con cada algoritmo, se destaca que todos ellos entregaron los mejores resultados empleando la distancia euclidiana y la VM1, donde esta última estaba compuesta por los datos estandarizados de la latitud, longitud, profundidad y magnitud. Esto permite concluir que la estandarización de los datos como las coordenadas geográficas aporta un valor importante para los algoritmos al momento de crear los grupos, a diferencia de las otras VM que no incluían las coordenadas geográficas o no tenían los datos estandarizados. Similar situación ocurre con la distancia euclidiana, la cual permitió que los algoritmos entreguen los mejores resultados en comparación a los experimentos donde se utilizando otras métricas de distancia.

Como trabajo a futuro se puede considerar estudiar el tipo de suelo y otras características más especializadas que tengan relación con topografía, geomorfología, entre otros que puedan complementar más información sobre los posibles PPRs identificados en este trabajo, así como también de los sectores que los rodean. Adicionalmente se puede considerar el punto de vista de habitabilidad (construcción, urbanización, etc.), desarrollo vial (carreteras, caminos, etc.) y factores climáticos presentes en el Norte Grande de Chile que puedan afectar a los PPRs identificados.

Finalmente, el trabajo se puede ampliar realizando experimentos con diferentes algoritmos basados en densidad ya que estos demostraron entregar los mejores resultados dada la capacidad para identificar grupos de datos no necesariamente convexos, con ello se podrían descubrir otras zonas geográficas adicionales que DBSCAN consideró como ruido. Otra posibilidad es someter los datos que fueron considerados como ruido a un proceso de *clustering* independiente para separar estos datos de los que ya fueron agrupados satisfactoriamente.

AGRADECIMIENTOS

Este trabajo ha sido parcialmente financiado por el proyecto postdoctoral (grant No. 74200094) de ANID, Chile.

REFERENCIAS

[1] E.J. Tarbuck y F.K. Lutgens, *Ciencias de la tierra: Una introducción a la geología física*, 8va ed. Madrid, España: Pearson Prentice Hall, 2005.

[2] M. Bejar-Pizarro *et al.*, "Asperities and barriers on the seismogenic zone in North Chile: state-of-the-art after the 2007 Mw 7.7 Tocopilla earthquake inferred by GPS and InSAR data," *Geophysical Journal International*, vol. 183, no. 1, pp. 390-406, Oct. 2010, doi: 10.1111/j.1365-246X.2010.04748.x.

[3] G. Valdebenito, D. Alvarado, C. Sandoval y V. Aguilar, "Terremoto de Iquique Mw = 8,2 - 01 abril 2014: daños observados y efectos de sitio en estructuras de albañilería", en *XI Congreso Chileno de Sismología e Ingeniería Sísmica*, vol. 221, Mar. 2015.

[4] J. Parraga-Alava, G.M. Garzón, R. Alcívar Cevallos and M. Inostroza-Ponta, "Unsupervised Pattern Recognition for Geographical Clustering of Seismic Events Post MW 7.8 Ecuador Earthquake," 2018 37th International Conference of the

Chilean Computer Science Society (SCCC), Santiago, Chile, 2018, pp. 1-8, doi: 10.1109/SCCC.2018.8705248.

[5] P. Novianti, D. Setyorini, and U. Rafflesia, "*K-Means* cluster analysis in earthquake epicenter *clustering*," *International Journal of Advances in Intelligent Informatics*, vol. 3, no. 2, pp. 81-89, Jul. 2017, doi: 10.26555/ijain.v3i2.100.

[6] M. Bariklana and A. Fauzan, "Implementation of the dbscan method for cluster mapping of earthquake spread location," *Barekeng*, vol. 17, no. 2, pp. 867-878, Jun. 2023, doi: 10.30598/barekengvol17iss2pp0867-0878.

[7] R. Kamat and R. Kamath, "Earthquake cluster analysis: K-means approach," *Journal of Chemical and Pharmaceutical Sciences*, vol. 10, no. 1, pp. 250-253, 2017.

[8] S. Khan, M. Waseem, S. Khalid, and D.P.N. Kontoni, "Fuzzy *clustering* analysis of HVSR data for seismic microzonation at lahore city," *Journal Shock and Vibration*, vol. 2022, Oct. 2022, doi: 10.1155/2022/3109609.

[9] P. Tapia G, W. Roldán L y C. Villacis, "Vulnerabilidad sísmica de las ciudades del norte de chile: Arica, Antofagasta y Copiapó", *VIII Jornadas Chilenas de Sismología e Ing. Antisísmica*, Valparaíso, Chile, 2002.

[10] R. Madariaga, "Sismicidad de Chile", *Física de la Tierra*, no. 10, pp. 221-258, 1998.

[11] C. Schröer, F. Kruse y J. Marx Gómez, "A systematic literature review on applying CRISP-DM process model," *Procedia Computer Science*, vol. 181, no. 2021, pp. 526-534, Jan. 2021, doi: 10.1016/j.procs.2021.01.199.

[12] J. Hernández, M.J. Ramírez y C. Ferri, *Introducción a la Minería de Datos*, 1ra ed. Madrid, España: Pearson Prentice Hall, 2004.

[13] J. Gironés, J. Casas, J. Minguillón y R. Caihuelas, *Minería de Datos: Modelos Y Algoritmos*, 1era ed. Barcelona, España: UOC, 2017.

[14] Y. Li, M. Dong, and J. Hua, "A gaussian mixture model to detect clusters embedded in feature subspace," *Communications in Information and Systems*, vol. 7, no. 4, pp. 337-352, May. 2007.

[15] M.J. Zaki and W. Meira, *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*, 2da ed. Cambridge, UK: Cambridge University Press, 2020.

[16] P.N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*, 1st ed. Boston, USA: Pearson Education, 2005.

[17] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "On *clustering* validation techniques," *Journal of Intelligent Information Systems*, vol. 17, no. 2, pp. 107-145, Dic. 2001, doi: 10.1023/A:1012801612483.

[18] P.J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster análisis," *Journal of Computational and Applied Mathematics*, vol. 20, no. 1, pp. 53-65, Nov. 1987, doi: 10.1016/0377-0427(87)90125-7.

[19] D.L. Davies and D.W. Bouldin, "A Cluster Separation Measure," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224-227, Apr. 1979, doi: 10.1109/TPAMI.1979.4766909.

[20] J. Dunn, "Well-separated clusters and optimal fuzzy partitions," *Journal of Cybernetics*, vol. 4, no. 1, pp. 95-104, Sep. 1974, doi: 10.1080/01969727408546059.

[21] C. Tadeusz and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1-27, Sep. 1974, doi: 10.1080/03610927408827101.

[22] K. Rojas-Jimenez, "Ciencia de datos para ciencias naturales," bookdown.org, 2022. [En línea]. Disponible en: https://bookdown.org/keilor_rojas/CienciaDatos/ (Accedido: 12 de marzo del 2023).

[23] M. Ester, H.P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, Portland, Oregon, 1996, pp. 226-231.

[24] D. Moulavi, P. Jaskowiak, R. Campello, A. Zimek, and J. Sander, "Density-Based *clustering* validation," in *Proceedings of the 2014 SIAM International Conference on Data Mining (SDM),* Apr. 2004, pp. 839-847, doi: 10.1137/1.9781611973440.96