UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERIA DE MINAS

DEFINICIÓN DE SETS ESTRUCTURALES MEDIANTE TÉCNICAS DE CLUSTERING

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL DE MINAS

SEBASTIAN GABRIEL REYES POBLETE

PROFESOR GUÍA:
FABIÁN SOTO FERNÁNDEZ

MIEMBROS DE LA COMISIÓN:
NADIA MERY GUERRERO
FELIPE NAVARRO VARGAS

SANTIAGO DE CHILE
2023

RESUMEN DE LA MEMORIA PARA OPTAR AL
TÍTULO DE: Ingeniero Civil de Minas
POR: Sebastian Gabriel Reyes Poblete
FECHA: 2023
PROFESOR GUÍA: Fabián Soto Fernández

DEFINICIÓN DE SETS ESTRUCTURALES MEDIANTE TÉCNICAS DE CLUSTERING

RESUMEN EJECUTIVO

El presente trabajo de investigación aborda el desafío de analizar y agrupar datos geotécnicos, específicamente las orientaciones de discontinuidades. La capacidad de comprender estas orientaciones es fundamental para la seguridad y eficiencia de proyectos en el área de ingeniería de minas y la geotecnia. En este contexto, se revisó una amplia gama de algoritmos de clustering para identificar patrones y estructuras en este tipo de datos.

El análisis se centró en la evaluación de varios algoritmos de clustering aplicados a un caso sintético y a dos casos de estudio reales. Entre los algoritmos considerados, se destacaron el *K-Means* y el *Spectral Clustering* por su efectividad y simplicidad en la especificación del número de clústeres.

Para definir el número de clústers se aplicaron métodos estadísticos, como el método del codo, el método de la silueta y el método *Gap Statistic*, donde se obtuvo que el número óptimo de clústeres era 6. Eso se condice con los resultados esperados del caso sintético.

Los casos de estudio demostraron que los algoritmos de clustering pueden identificar patrones significativos en las orientaciones de las discontinuidades geotécnicas, a pesar de las diferencias en las distribuciones de datos y la complejidad geológica.

ABSTRACT OF THE THESIS TO OBTAIN THE GRADE OF: Mining Engineer
BY: Sebastian Gabriel Reyes Poblete
DATE: 2023
THESIS ADVISOR: Fabián Soto Fernández

DEFINITION OF STRUCTURAL SETS THROUGH CLUSTERING TECHNIQUES
EXECUTIVE SUMMARY

This research work addresses the challenge of analyzing and clustering geotechnical data, specifically the orientations of discontinuities. The ability to understand these orientations is fundamental for the safety and efficiency of projects in the fields of mining engineering and geotechnics. In this context, a wide range of clustering algorithms was reviewed to identify patterns and structures in this type of data.

The analysis focused on evaluating several clustering algorithms applied to a synthetic case and two real case studies. Among the algorithms considered, K-Means and Spectral Clustering stood out for their effectiveness and simplicity in specifying the number of clusters.

To determine the number of clusters, statistical methods were applied, such as the elbow method, the silhouette method, and the Gap Statistic, which consistently indicated that the optimal number of clusters was 6. This aligns with the expected results from the synthetic case.

The case studies demonstrated that clustering algorithms can identify significant patterns in the orientations of geotechnical discontinuities, despite differences in data distributions and geological complexity.

TABLA DE CONTENIDOS

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

1 INTRODUCCIÓN

La industria minera durante el siglo pasado tuvo como principal objetivo el cumplir con las metas de producción. Esto se ha visto modificado en la última década al incorporar nuevos desafíos para el desarrollo de la industria. Dentro de estos desafíos nos encontramos con una legislación más exigente, con comunidades que demandan una minería más sustentable, con desafíos técnicos (como la disminución de las leyes de los yacimientos, entre otros), y con personas que priorizan la seguridad antes de realizar una labor en la faena, personas que al final del día son las que mantienen el negocio minero funcionando.

En este último punto, la seguridad se ha transformado actualmente en una prioridad para las faenas mineras, y en las cuales ya es considerada como un valor, logrando agregar valor a cada operación que se lleve a cabo dentro de ellas y así mantener la continuidad operacional. Por esto es importante entregar datos con una calidad mayor en la caracterización del macizo rocoso, y así disminuir los riesgos y accidentes que pueden ocurrir en la mina. Mientras mejor es la calidad de la información disponible, mejores serán las decisiones tomadas con dicha información.

Para establecer la estabilidad de un macizo rocoso existen distintas definiciones y estudios al respecto, además, se consideran diversos factores. Un factor común a todas las definiciones es la orientación de las estructuras, y más aún, de las estructuras principales. La obtención de estas orientaciones se realiza analizando la información estructural mediante proyecciones estereográficas en la cual cada estructura se representa por un punto o polo. Luego, mediante técnicas estadísticas se agrupan estos polos en agrupaciones para definir así los sets estructurales principales.

Con lo anterior en mente, la motivación principal este trabajo de memoria es mejorar la calidad de la información geotécnica asociada a la definición de sets estructurales. Esta definición de los sets en principio se realiza en softwares dónde la agrupación de polos se realiza manualmente. En los últimos años esta tendencia ha incorporado algoritmos de *clustering* determinísticos para la definición de estos sets.

Los algoritmos considerados para evaluar su desempeño corresponden a los siguientes: algunos implementados actualmente en softwares comerciales, tales como *K-Means*, y otros que a la fecha no han sido utilizados para tal trabajo, tales como DBSCAN, *Mean-Shift*, entre otros.

En el primer capítulo se contextualiza el trabajo. En el capítulo 2 se presentan los objetivos y los alcances. En el capítulo 3 se presentan las bases teóricas de clustering y geología estructural que permitirán tener un lenguaje común en las secciones posteriores. En el capítulo 4 se presenta la revisión bibliográfica donde se implementan algoritmos de clustering en la definición de sets estructurales. En el capítulo 5 se presenta la metodología de agrupamiento. En el capítulo 6 se presentan los resultados para los algoritmos seleccionados en un caso sintético. En el capítulo 7 se presenta los resultados para 2 casos de estudio con datos reales. En el capítulo 8 se discuten y analizan los resultados obtenidos en los capítulos previos. Finalmente, en el capítulo 9 se presentan las conclusiones y recomendaciones del trabajo.

2 OBJETIVOS Y ALCANCES

2.1 Objetivo general

El objetivo general de este trabajo es evaluar el desempeño de una variedad de algoritmos de clustering en la identificación de sets estructurales en redes estereográficas, con el fin de mejorar la calidad y eficiencia de la caracterización geotécnica en el contexto de la geología estructural aplicada a la industria minera.

2.2 Objetivos específicos

Los objetivos específicos del presente trabajo consisten en:

1. Seleccionar algoritmos de clustering no supervisados, como *K-Means, Spectral Clustering, Affinity Propagation, DBSCAN, Mean-Shift* para el agrupamiento de sets estructurales
2. Desarrollar rutinas de programación en Python para la ejecución de los algoritmos seleccionados con el fin de evaluar su desempeño.
3. Evaluar el desempeño de los algoritmos de clustering ejecutados para los diagramas de contorno mediante validaciones visuales y métricas.
4. Relacionar parámetros de la inicialización de los algoritmos con el número de sets estructurales obtenidos después de su aplicación, utilizando tanto un caso de prueba sintético como dos casos de estudio reales.

2.3 Alcances

Los alcances del presente trabajo de título son los siguientes:

1. Se llega hasta la identificación de sets estructurales. La definición posterior de dominios estructurales incorporando otras variables queda para trabajos posteriores.
2. La correlación geoespacial de los datos estructurales no es considerada en la identificación de los sets estructurales.
3. Los algoritmos seleccionados y utilizados para evaluar su desempeño en la definición de sets estructurales serán de obtenidos de librerías en *scikit-learn.*
4. La representación gráfica en redes estereográficas será realizada a través de la librería *mplstereonet.*
5. La programación corresponderá a una rutina para ejecutar las librerías mencionadas en Python.

3 MARCO TEÓRICO

A continuación, se presentan las bases teóricas utilizadas para la obtención de resultados.
Se presentan las nociones de *clustering* y los algoritmos para realizarlo. También se presentan fundamentos de geología estructural.

3.1 Clustering

El *clustering* o agrupamiento corresponde a juntar elementos similares de un conjunto en grupos, según un criterio previamente definido por el usuario [7].

Existen diversos enfoques para el clustering, cada uno de los cuales se adapta a diferentes tipos de datos y escenarios. Algunos de los tipos de clustering más comunes son:

1. *Clustering* basado en conectividad: Los elementos se agrupan si están conectados de alguna manera, como en una red o grafo.
2. *Clustering* basado en centroides: Los elementos se agrupan alrededor de un centroide (como *K-Means*).
3. *Clustering* basado en distribuciones: Los elementos se agrupan en función de sus similitudes en su distribución de probabilidad.
4. *Clustering* basado en densidad: Los elementos se agrupan identificando regiones de alta densidad de puntos como clústeres (como *DBSCAN*).

Cada uno de estos enfoques mencionados tiene sus propias ventajas y desventajas y se adapta a diferentes tipos de datos y problemas.

3.1.1 Algoritmos de *Clustering*

Los algoritmos de *clustering* corresponden a la secuencia de pasos sucesivos para realizar el agrupamiento. Existen múltiples tipos de algoritmos reportados en la literatura [6].

Se debe mencionar que el algoritmo más apropiado para un problema en particular suele ser escogido experimentalmente, salvo que exista alguna razón topológica para preferir un algoritmo sobre otro. Además, es fundamental tener en cuenta que un algoritmo diseñado para un modelo específico de datos puede no funcionar de manera eficaz en un conjunto de datos que sigue un modelo diferente [6].

Los algoritmos de *clustering* seleccionados para este estudio corresponden a los siguientes (revisados en las siguientes subsecciones):

1. *K-Means*
2. *Spectral Clustering*
3. *Affinity Propagation*
4. *DBSCAN (Density-Based Spatial Clustering of Applications with Noise)*
5. *Mean-Shift*

3.1.1.1 K-Means

Es un algoritmo de aprendizaje sin supervisión que resuelve el problema de agrupamiento. El procedimiento permite clasificar un determinado conjunto de datos de manera simple y fácil, a

través de un número K de *clusters* o grupos, donde el número K de grupos debe estar fijado a priori.

La idea principal es definir K centroides, uno para cada grupo. Estos centroides deben ser colocados de una manera astuta debido a la ubicación diferente causa resultado diferente. Por lo tanto, la recomendación es colocarlos lo más lejos uno del otro. El siguiente paso consiste en tomar cada punto perteneciente a un conjunto de datos dado y asociarlo al centroide más cercano. Una vez hecho lo anterior con todos los datos, se completa el primer paso y se hace una agrupación temprana. A partir de este punto es necesario volver a calcular K nuevos centroides análogamente a la etapa anterior. Después de tener estos K nuevos centroides, se deben volver a asociar todos los datos a su nuevo centroide más cercanos. Se repiten los dos pasos anteriores hasta que ninguno de los K centroides se modifiquen (según un parámetro épsilon de tolerancia), donde se dice que el algoritmo convergió [3].

En resumen, los principales pasos del algoritmo son los siguientes:

1. Colocar los puntos K en el espacio representado por los objetos que se están agrupando. Estos puntos representan los centroides del grupo inicial.
2. Asignar cada dato al grupo que tenga el centroide más cercano.
3. Cuando se hayan asignado todos los datos, volver a calcular las posiciones de los K centroides.
4. Repetir los pasos 2 y 3 hasta que los centroides ya no se muevan. Esto produce una separación de los objetos en grupos en los que se puede calcular la métrica a minimizar.

Los pasos del algoritmo se ejemplifican en la Figura 3-1.

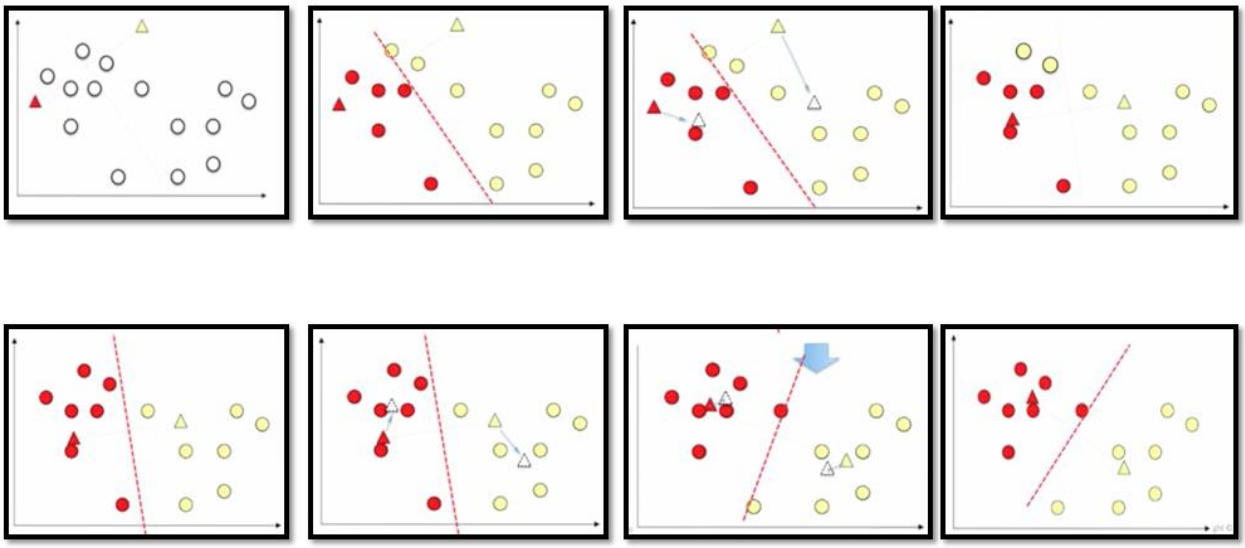

Figura 3-1 Ejemplo de pasos del algoritmo *K-Means*

Finalmente, este algoritmo tiene como objetivo minimizar una función objetivo, que en este caso corresponde a una función de error cuadrático. La función objetivo es la siguiente:

$$\sum_{i=0}^{n} \min_{u_j \in C} \left(\|x_j - u_i\|^2 \right)$$

Ecuación 1: Función objetivo a minimizar en K-Means.

Donde:
n: número de muestras
x_j: muestra j
u_j: media de las muestras en el clúster
C: conjunto de clústeres disjuntos

A pesar de que, se puede probar que el algoritmo converge siempre, *K-Means* no encuentra necesariamente la configuración óptima, correspondiente a la función objetivo global mínima. El algoritmo es significativamente sensible a los centros de agrupación seleccionados al azar inicial. Se puede ejecutar varias veces para reducir este efecto [3].

A continuación, en la Figura 3-2, se presenta un ejemplo del resultado que entrega el algoritmo *K-Means*, donde vemos la partición del espacio en *clusters* cuyos centroides corresponden a las X en color blanco.

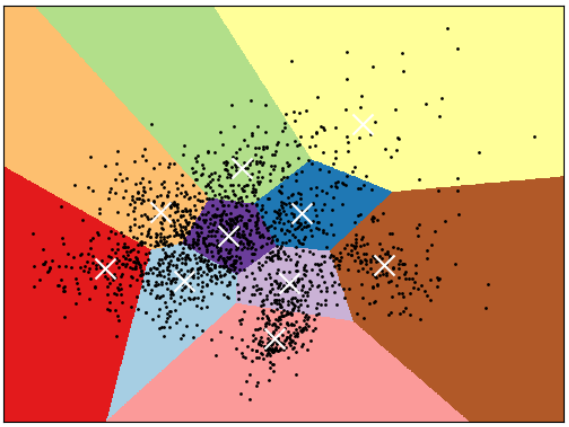

Figura 3-2: Ejemplo de resultado entregado por el algoritmo *K-Means*. Las X blancas representan los centroides [1].

Las ventajas del método:

1. Es relativamente fácil de implementar y comprender.
2. Es eficiente en términos computacionales, lo que lo hace adecuado para conjuntos de datos grandes.

Las desventajas del método [15]:

1. Es sensible a la inicialización de los centroides, lo que significa que puede converger a mínimos locales en lugar del óptimo global.
2. Es sensible al número de clústeres escogido.
3. Es sensible a datos *outliers*.
4. Asume que los clústeres son esféricos y de igual tamaño.

3.1.1.2 Spectral Clustering

El objetivo del *spectral clustering* es agrupar los datos que están conectados, pero no necesariamente agrupados dentro de límites convexos.

Dado un conjunto enumerado de puntos de datos, la matriz de similitud se puede definir como una matriz simétrica A, donde la componente A_{ij} representa una medida de la similitud entre los puntos de datos con índices i y j, denominada matriz de afinidad. En la Figura 3-3 se presenta un esquema de dicha matriz.

Una vez definida esta matriz de afinidad, la agrupación en *clusters* se reemplaza por un problema de división de grafos, donde los componentes del grafo conectados se interpretan como *clusters*. El gráfico debe estar particionado de manera que los bordes que conectan diferentes conglomerados tengan pesos bajos, y los bordes dentro del mismo conglomerado deben tener valores altos [16].

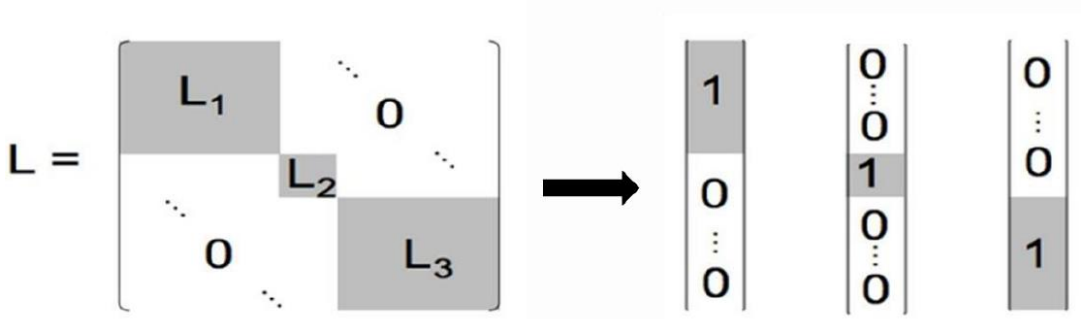

Figura 3-3 Representación esquemática de la matriz de afinidad. A esta matriz se le calculan los vectores propios para la división de los grafos.

A diferencia del *K-Means*, este método no asume la forma o tamaño de los clústeres, lo que le permite identificar agrupaciones de datos no convexas y de tamaños variables de forma eficaz. Sin embargo, se debe tener en cuenta su sensibilidad a la selección de parámetros críticos, como el número de clústeres. También se debe tener en cuenta que su capacidad de escalabilidad puede verse limitada al aplicarse a conjuntos de datos muy grandes (por el cálculo y la diagonalización de la matriz de afinidad).

3.1.1.3 Affinity Propagation

El algoritmo *affinity propagation* crea clústeres enviando mensajes entre pares de muestras hasta la convergencia. Un set de datos se describe utilizando un pequeño número de ejemplares, que se identifican como los más representativos de otras muestras. Los mensajes enviados entre pares representan la idoneidad para que una muestra sea el ejemplo de la otra, que se actualiza en respuesta a los valores de otros pares. Esta actualización ocurre de forma iterativa hasta la convergencia, momento en el que se eligen los ejemplos finales y, por lo tanto, se da la agrupación final.

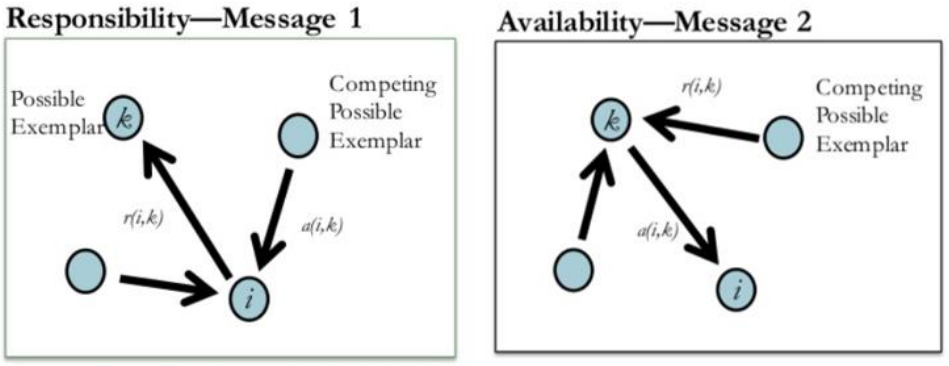

Figura 3-4 Representación del envío de mensajes en los pares de datos para el Affinity Propagation.

Los pasos del método son los siguientes:

1. Se asigna un valor de preferencia a cada muestra, que refleja cuán deseable es que una muestra sea un ejemplo representativo de otras.
2. Se calculan las similitudes entre pares de muestras utilizando métricas de similitud, como la distancia euclidiana.
3. Se envían mensajes entre pares de muestras, donde cada mensaje representa la idoneidad de una muestra para ser el ejemplo de la otra.
4. Se actualizan dichos mensajes iterativamente en respuesta a los valores de otros pares.
5. Se determina la convergencia cuando ya no se producen cambios en los mensajes.

Las ventajas del método:

1. No se requiere especificar el número de clústeres.
2. El algoritmo es capaz de identificar clústeres de diferentes tamaños.
3. Permite la identificación de ejemplos representativos.
4. No es tan sensible a la elección inicial de los puntos de referencia y puede converger de manera efectiva.

Las desventajas del método:

1. Puede ser computacionalmente costoso en conjuntos de datos grandes (pues implica el cálculo iterativo de mensajes entre todas las parejas de datos).
2. Podría generar un número excesivo de clústeres.

3.1.1.4 Agglomerative Clustering

Los algoritmos de agrupamiento jerárquico se dividen en 2 categorías: de arriba hacia abajo o de abajo hacia arriba. Los algoritmos ascendentes tratan cada punto de datos como un solo grupo al principio y luego combinan (o aglomeran) sucesivamente pares de grupos hasta que todos los grupos se hayan fusionado en un solo grupo que contiene todos los puntos de datos. El agrupamiento jerárquico ascendente se denomina agrupamiento jerárquico aglomerativo. Esta jerarquía de grupos se representa como un árbol (o dendrograma). La raíz del árbol es el grupo

único que reúne todas las muestras, siendo las hojas los grupos con una sola muestra. Para la elección de *clusters* se pueden realizar cortes en las ramas del árbol.

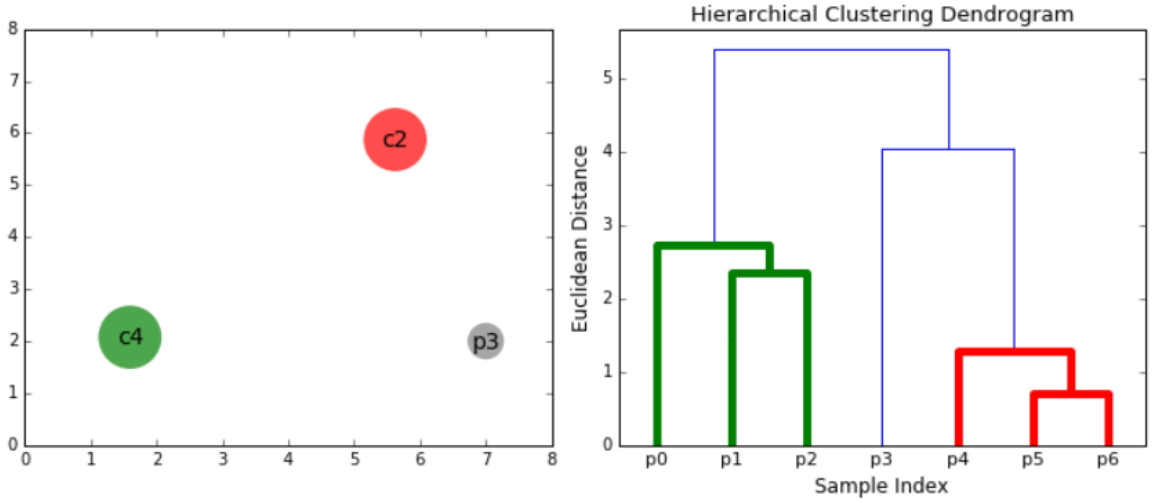

Figura 3-5 Ejemplo de *clusters* mediante *Agglomerative Clustering* (izquierda) y su dendrograma respectivo (derecha).

El agrupamiento jerárquico ascendente (Aglomerativo) considera cada dato como un clúster individual y luego los fusiona (aglomera) sucesivamente en pares de clústeres hasta que todos los datos se agrupen en un solo clúster. El resultado de este proceso es una jerarquía de clústeres que se representa típicamente a través de un dendrograma, donde la raíz del árbol representa un clúster que contiene todos los datos, y las hojas representan clústeres individuales que contienen un solo dato. El dendrograma permite seleccionar clústeres en función de la altura de corte en las ramas del árbol. Cuanto más bajo sea el corte realizado, más clústeres se obtendrán, pero a medida que se aumente la altura de corte, se obtendrá un número menor de clústeres.

El agrupamiento jerárquico descendente (divisivo) considera un clúster que contiene todos los datos y luego los divide repetidamente en subclústeres más pequeños. A diferencia del ascendente, el resultado de este proceso es una jerarquía de clústeres que comienza con un clúster grande en la raíz y se ramifica en subclústeres más pequeños a medida que desciende en el árbol.

Las ventajas del agrupamiento jerárquico ascendente:

1. La jerarquía de clústeres se representa de manera intuitiva a través de un dendrograma (esto facilita la interpretación de los resultados).
2. No es necesario especificar previamente el número de clústeres.
3. Se puede elegir la altura de corte en el dendrograma (esto proporciona flexibilidad en la identificación de clústeres).
4. Puede ser robusto ante datos ruidosos (debido a la fusión de clústeres).

Las desventajas del agrupamiento jerárquico ascendente:

1. Puede ser computacionalmente costoso para conjuntos de datos grandes.
2. No es posible deshacer esa fusión en el proceso aglomerativo.
3. Diferentes inicializaciones pueden llevar a jerarquías de clústeres diferentes.
4. La interpretación de los resultados puede ser subjetiva.

3.1.1.5 DBSCAN

Es un algoritmo de agrupamiento basado en la densidad de los datos, es decir, dado un conjunto de puntos en algún espacio, agrupa los puntos que están estrechamente agrupados (puntos con muchos vecinos cercanos), marcando como puntos atípicos los puntos que se encuentran solos en las regiones de baja densidad (las más cercanas los vecinos están muy lejos).

Se deben identificar tres puntos importantes: el núcleo, el borde y el ruido; y dos parámetros para la inicialización: el número mínimo de puntos y el radio de búsqueda [20].

- ✓ Núcleo (Core): Corresponde a puntos centrales que tienen al menos un número mínimo de puntos (definido por el parámetro "número mínimo de puntos") en su vecindario dentro de un radio específico (definido por el parámetro "radio de búsqueda"). Los puntos en el interior del área densa formada por núcleos son considerados como parte de un clúster.
- ✓ Borde (Border): Corresponde a puntos que no son núcleos, pero se encuentran en el vecindario de un núcleo (dentro del radio de búsqueda). Estos puntos se asignan al mismo clúster que su núcleo más cercano.
- ✓ Ruido (Noise): Corresponde a puntos que no son núcleos ni puntos de borde. Es decir, no pertenecen a ningún clúster. Estos son los puntos aislados que no pueden agruparse en clústeres densos.
- ✓ Número mínimo de puntos: Este parámetro determina cuántos puntos deben estar dentro del radio de búsqueda para que un punto se considere un núcleo.
- ✓ Radio de búsqueda: Este parámetro corresponde a la distancia máxima en la que se busca a otros puntos dentro de un núcleo para determinar si están lo suficientemente cerca (para formar parte del mismo clúster).

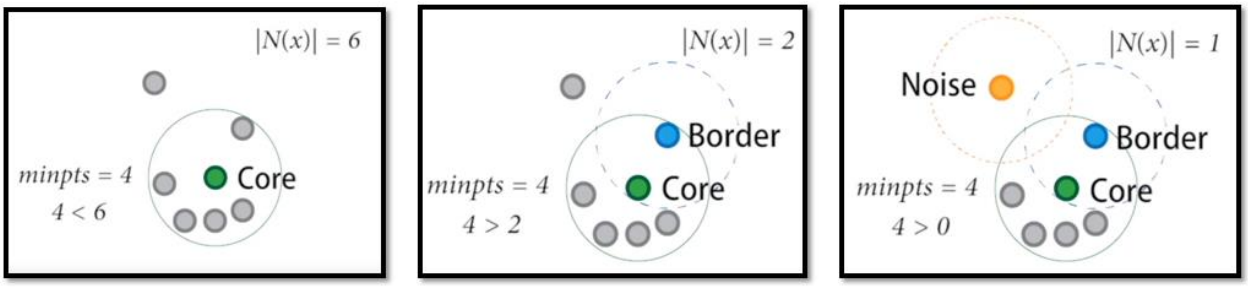

Figura 3-6 Representación esquemática del concepto de núcleo, borde y ruido para DBSCAN.

Los pasos del algoritmo son los siguientes:

1. Se selecciona un punto inicial no visitado anteriormente.
2. Se calcula la vecindad del punto seleccionado dentro de un radio específico (definido por el radio de búsqueda). Todos los puntos que caen dentro de este radio se consideran vecinos del punto seleccionado.

3. Se verifica si el número de vecinos dentro del radio es igual o mayor que el número mínimo de puntos (definido por el número mínimo de puntos). Si se cumple esta condición, el punto seleccionado se clasifica como un núcleo.
4. Se inicia la formación de un nuevo clúster con el punto núcleo.
5. Se revisan todos los vecinos de este núcleo y se comprueba si cada vecino es también un núcleo. Si es así, se agregan sus vecinos al clúster.
6. El proceso se repite recursivamente hasta que no se puedan agregar más puntos al clúster.
7. Una vez que se haya explorado completamente el clúster actual, el algoritmo vuelve al paso 1 para seleccionar otro punto no visitado.
8. Luego de que todos los puntos hayan sido clasificados en clústeres, se clasifican los puntos que no son núcleos, pero están en el vecindario de un núcleo como puntos de borde. Estos puntos son asignados al mismo clúster que su núcleo más cercano.
9. Finalmente, los puntos que no se han asignado a ningún clúster se consideran ruido.

Las ventajas del método:

1. Puede identificar clústeres de diferentes formas y tamaños.
2. Puede detectar puntos atípicos (ruido).
3. No requiere especificar el número de clústeres.
4. Es eficaz en la identificación de clústeres basados en densidad.
5. Es menos dependiente de la configuración inicial.

Las desventajas del método [15]:

1. Es sensible a la elección de los parámetros
2. Puede ser computacionalmente costoso, especialmente en conjuntos de datos grandes.
3. No es adecuado para bases de datos con densidades muy variables (puede tener dificultades para encontrar una estructura de clústeres coherente).
4. No es adecuado para bases de alta dimensionalidad.

3.1.1.6 Mean-Shift

El algoritmo de Mean Shift tiene como objetivo descubrir puntos en una densidad uniforme de muestras. Es un algoritmo basado en centroide, que funciona actualizando los candidatos para que los centroides sean la media de los puntos dentro de una región determinada. Estos candidatos luego se filtran en una etapa de post procesamiento para eliminar duplicados para formar el conjunto final de centroides [5].

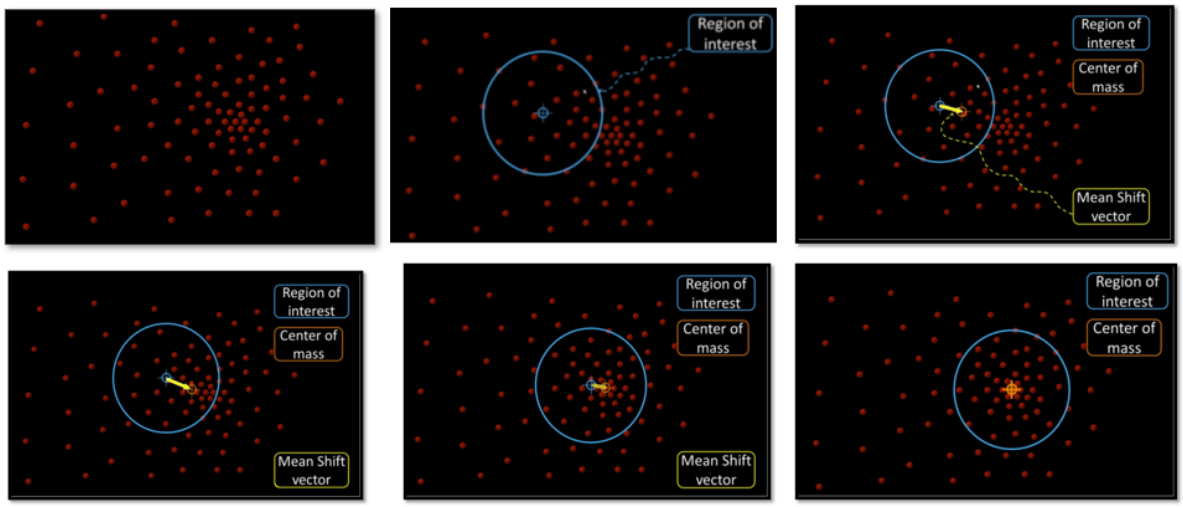

Figura 3-7 Iteraciones del algoritmo Mean-Shift.

Las ventajas del método:

1. Puede identificar clústeres de diferentes formas y tamaños.
2. No se requiere especificar el número de clústeres previamente.
3. Es efectivo para identificar clústeres basados en densidad.
4. Es adecuado para clústeres con densidades variables.

Las desventajas del método:

1. Es sensible al ancho de banda (bandwidth).
2. Puede ser computacionalmente costoso, especialmente en conjuntos de datos grandes.
3. No es adecuado para datos de alta dimensionalidad.
4. Puede ser sensible a valores atípicos en el conjunto de datos.

3.1.2 Resumen de los parámetros necesarios para los algoritmos

Para efectos prácticos, lo principal que se requiere conocer corresponde a los parámetros de entrada para la inicialización de cada algoritmo.

Tabla 3-1 Resumen de parámetros de iniciación para los algoritmos de *clustering* seleccionados.

Algoritmo	Parámetros
K-Means	N° de clústers
Affinity Propagation	*Damping*
Spectral Clustering	N° de clústers Tipo de unión *(Ward, average)*
Agglomerative Clustering	N° de clústers

DBSCAN	Radio de búsqueda, número mínimo de puntos
Mean-Shift	No es necesario *(bandwidth)*

3.2 Métricas para evaluar el desempeño de los algoritmos

3.2.1 Elección del número de clústeres

Existen distintas técnicas para seleccionar el número de clústeres, tales como el método del codo, el método de la silueta, el método *gap statistic*, el método visual, entre otros.

3.2.1.1 Método del codo *(Elbow Method)*

El método del codo es una técnica utilizada para determinar el número óptimo de clústeres en un algoritmo de agrupamiento, como K-Means. Se basa en el cálculo de la varianza intra-clúster para diferentes valores de k, donde k representa el número de clústeres. La idea es encontrar un equilibrio entre tener suficientes clústeres para representar estructuras significativas en los datos y evitar un exceso de clústeres que pueda conducir a una sobre segmentación. Se ejecuta el algoritmo de agrupamiento para un rango de valores de k, y se calcula la suma de las varianzas intra-clúster. Luego, se grafican estos valores en función de k. El punto en el que la disminución en la varianza intra-clúster se estabiliza y comienza a aplanarse se considera el número óptimo de clústeres, ya que representa un equilibrio entre la compacidad de los clústeres y la representación de estructuras significativas en los datos [12].

3.2.1.2 Método de la silueta *(Silhouette Method)*

El Método de la Silueta (Silhouette Method) es una técnica que se utiliza para evaluar el número óptimo de clústeres en algoritmos de agrupamiento, como K-Means. Este enfoque se basa en la medida de la silueta, que evalúa cuán bien se han agrupado los puntos de datos en clústeres. Para determinar el número óptimo de clústeres, se calcula la silueta promedio para diferentes valores de k (el número de clústeres). Cuanto mayor sea el valor de silueta, mejor será la calidad del agrupamiento. El número de clústeres que produce el valor de silueta promedio más alto se considera el número óptimo. El Método de la Silueta es útil para encontrar un equilibrio entre la compacidad de los clústeres y la separación entre ellos, lo que conduce a una representación eficaz de las estructuras de los datos en clústeres [4].

3.2.1.3 Método *Gap Statistic*

El Método Gap Statistic es una técnica utilizada para determinar el número óptimo de clústeres en algoritmos de agrupamiento. Este método compara la dispersión intra-clúster de los datos reales con la dispersión esperada en datos generados al azar. El objetivo es identificar el número de clústeres que ofrece una mayor cohesión de datos en comparación con una agrupación aleatoria. Se calcula una métrica Gap para diferentes valores de k (número de clústeres), y el número óptimo de clústeres se elige cuando la métrica Gap alcanza su valor máximo. El Método Gap Statistic proporciona una forma objetiva de seleccionar el número adecuado de clústeres al comparar el agrupamiento real con lo que podría esperarse por azar, lo que ayuda a evitar la sobresegmentación o la subsegmentación en el análisis de datos.

3.2.2 Evaluación del desempeño

Existen distintas métricas que se utilizan para evaluar el desempeño de algoritmos de *clustering*, tales como:

1. Índice de Silueta *(Silhouette Score):* Esta métrica evalúa la cohesión y la separación de los clústeres. Cuanto más alto sea el valor de silueta, mejor será la calidad del agrupamiento. Los valores van de -1 a 1, donde un valor cercano a 1 indica que los puntos están bien agrupados, un valor cercano a 0 indica superposición de clústeres y un valor cercano a -1 sugiere que los puntos se agrupan incorrectamente.
2. Índice Calinski-Harabasz *(Variance Ratio Criterion):* Esta métrica mide la relación entre la dispersión intra-clúster y la dispersión inter-clúster. Valores más altos indican una mejor separación entre clústeres.
3. Índice Davies-Bouldin: Esta métrica cuantifica la similitud promedio entre cada clúster y su clúster más similar. Valores más bajos indican una mejor partición de clústeres.
4. Puntuación de Dunn: Evalúa la relación entre la distancia mínima entre clústeres y la distancia máxima dentro de los clústeres. Un valor más alto indica una mejor separación entre clústeres.
5. Índice Rand Ajustado *(Adjusted Rand Index, ARI):* Esta métrica mide la similitud entre el agrupamiento obtenido y una agrupación de referencia (por ejemplo, etiquetas verdaderas). Los valores oscilan entre -1 y 1, donde 1 indica una concordancia perfecta.

3.3 Elementos de Geología Estructural

3.3.1 Dominio estructural

Si bien no existe una definición formal podemos definir un dominio estructural como una porción de terreno; un volumen delimitado geográficamente con una longitud, latitud y altura; un conjunto de litologías, relacionadas cinemáticamente y delimitadas por un conjunto de estructuras mayores (fallas) con características estructurales similares. Es decir, podemos definirlo con un conjunto de familias estructurales (tipos de estructuras) preferenciales principales y/o secundarias.
Usualmente se utiliza la distribución espacial de fracturas o sets de discontinuidades para identificar regiones o dominios geológicos estructurales que exhiben conjuntos de fracturas similares. Un ejemplo de la definición de dominio estructural se presenta en la Figura 3-8.

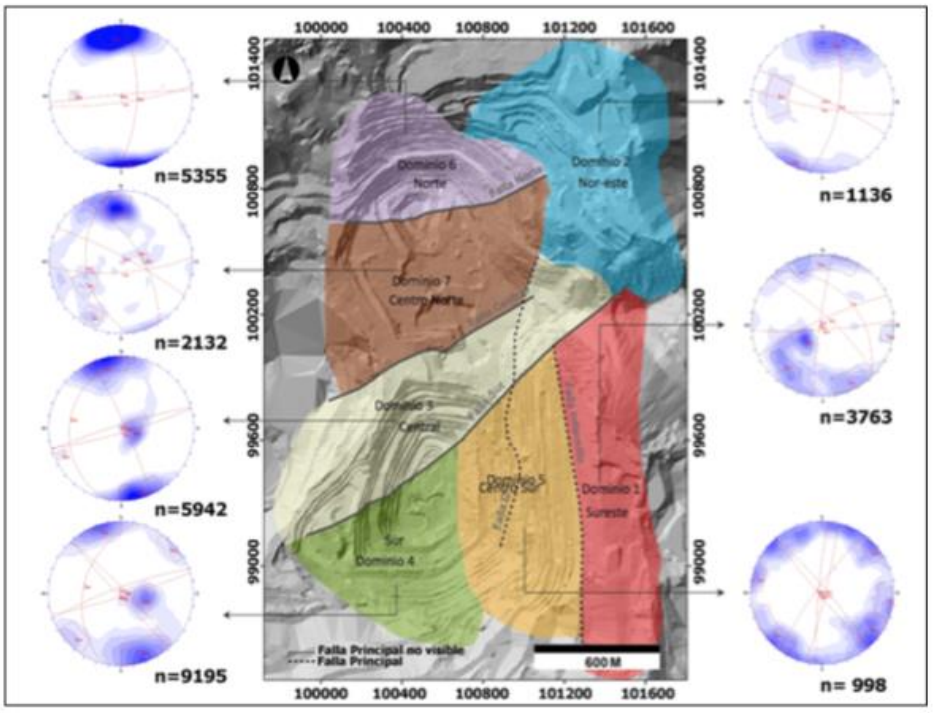

Figura 3-8 Dominios estructurales de la Mina Los Bronces. En la figura se observan los siete dominios estructurales definidos en el Modelo Estructural 2013 junto al diagrama de contorno de polos respectivo a cada dominio. Además, se aprecian las Fallas Principales, controladoras del límite del Dominio 3, 4, 5, 6 y 7; y la cantidad de datos (n) utilizados en la definición de cada uno de estos [2].

3.3.2 Set de discontinuidades

Una fractura corresponde a una superficie donde el material ha perdido cohesión. Una fractura que muestra un pequeño desplazamiento normal a su superficie y muy pequeño desplazamiento paralelo a la superficie de fractura se le denomina diaclasa o discontinuidad. Si se tiene un conjunto de estas discontinuidades que poseen una geometría plana, una orientación regular y paralela, y un espaciamiento regular se les denominan diaclasas sistemáticas.
Finalmente, un set de discontinuidades corresponde a un conjunto de diaclasas vecinas que tienen una geometría común [14].

3.3.3 Rumbo y manteo de una discontinuidad

La dirección y sentido de una estructura geológica (fractura, diaclasa, falla, etc) se puede representar mediante una línea, la cual está caracterizada por 2 ángulos: el azimut (o rumbo) y el buzamiento (o manteo). Entonces, se define el azimut de una línea como la dirección respecto al norte, mientras que, el buzamiento de una línea, como el ángulo respecto a un plano horizontal. En la Figura 3-9 se representan el rumbo y el manteo [18].

Figura 3-9: Rumbo y manteo.

3.3.4 Redes estereográficas

Para representar una discontinuidad, caracterizada por su rumbo y manteo, se utilizan las redes estereográficas. Una red estereográfica es una representación en 2 dimensiones de una esfera en la que es posible ubicar estructuras planares como fallas, fracturas, diaclasas, etc. Para esto basta con tener una medición del rumbo y manteo de la estructura. En la figura 3 se muestran 2 ejemplos de redes estereográficas, a la izquierda la red de Wulff y a la derecha la red de Schmidt [18].

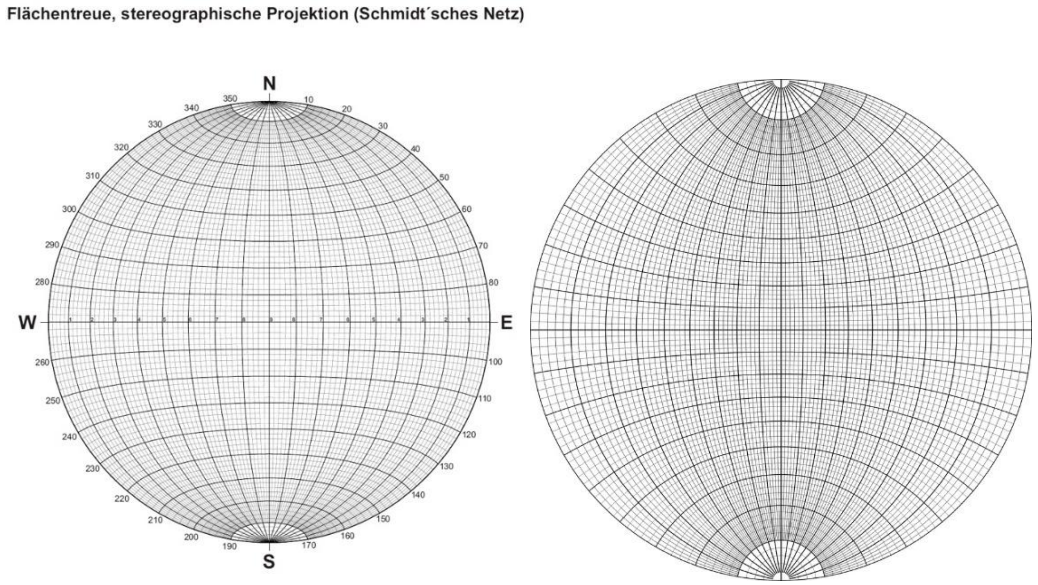

Figura 3-10: Red de Schmidt (izquierda) y Red de Wulff (derecha).

En geología estructural se usa la red de Schmidt, proyectando en el hemisferio inferior de la esfera. Se evita una concentración muy grande de puntos en el centro de la red, como ocurriría con una red de Wulff.
Una estructura planar se representa por una línea curva sobre la red, como se observa en la figura 4. A cada curva se le puede asociar un punto que la caracteriza de manera única, denominado polo. [18]

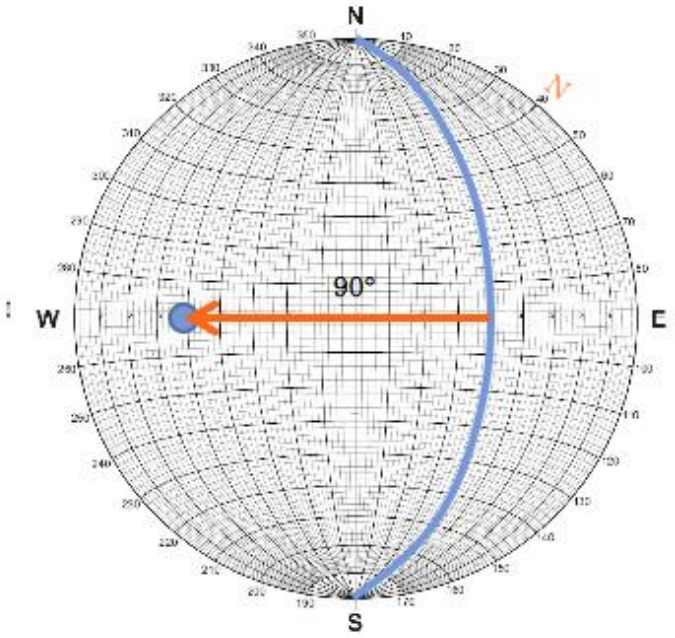

Figura 3-11: Representación de una estructura planar y su polo respectivo.

3.3.5 Diagrama de contorno

Un diagrama de contorno es una representación gráfica que muestra la distribución espacial de las orientaciones de las discontinuidades en un macizo rocoso que permite analizar las concentraciones máximas de polos. Se usa para visualizar el agrupamiento de datos de orientaciones que no se evidencian de inmediato a partir de un Gráfico de Polos o un Gráfico de dispersión. Los contornos representan concentraciones estadísticas de polos, calculadas mediante un método de distribución, que puede ser Fisher o Schmidt.

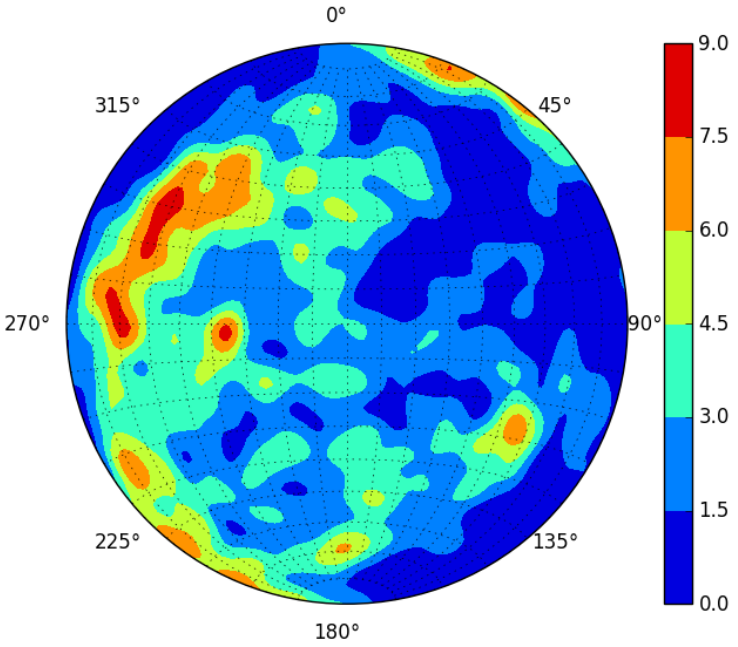

Figura 3-12 Ejemplo de diagrama de contorno.

Existen dos gráficos de contorno: uno no ponderado y otro ponderado con la ponderación de Terzagui. Al aplicar la corrección de Terzaghi al conjunto de datos usando el factor de ponderación, se puede obtener un diagrama de contornos “ponderado”, que representa mejor la concentración real de discontinuidades [22].

4 ESTADO DEL ARTE

En el presente capítulo se presenta la revisión de estudios realizados en el área de *clustering* de discontinuidades que han permitido el desarrollo de la metodología actual utilizada en este trabajo.

Por ejemplo, uno de los algoritmos más utilizado corresponde al *k-means* el cual fue nombrado por primera vez por James MacQueen en 1967 [13]. El desarrollo de estos algoritmos derivó en la evaluación y posibles áreas de aplicación de estos. En 1971, encontramos el estudio realizado por Rand sobre la evaluación de métodos de *clustering* [17].

4.1 Criterios Objetivos para la Evaluación de Métodos de *Clustering*

El problema principal consiste en determinar, dado un número de elementos, cómo seleccionar aquellos que están más cerca entre sí que con el resto de los objetos. Una vez resuelto este problema, es necesario comparar las soluciones. En general, existen dos formas de comparar los métodos de agrupación [17]:

- ✓ Usabilidad: Esto se refiere a cuán fácil es usar el método, a menudo relacionado con su capacidad de implementación computacional.
- ✓ Rendimiento: Esto implica cuán bien se desempeña el método.

Para decidir cuál método es mejor, se deben considerar las siguientes preguntas:

1. ¿Qué tan bien el método permite recuperar clústeres naturales? Esto evalúa la capacidad del método para identificar clústeres que son inherentes a los datos.
2. ¿Qué tan sensible es el método a perturbaciones en los datos? Se trata de cómo responde el método a cambios o perturbaciones en los datos.
3. ¿Qué tan sensible es el método ante datos faltantes? Esto examina cómo maneja el método situaciones en las que faltan algunos puntos de datos.
4. Si aplicamos dos métodos diferentes a los mismos datos, ¿obtenemos los mismos resultados? Esto verifica la consistencia y concordancia entre los métodos.

En resumen, la evaluación de métodos de agrupación debe considerar su capacidad para recuperar clústeres naturales, su sensibilidad a cambios en los datos, su manejo de datos faltantes y la consistencia de los resultados en comparación con otros métodos.

A continuación, se revisará lo realizado en cuanto a la aplicación de Métodos de Clustering en la agrupación de sets estructurales.

4.2 *Spectral Clustering* para identificación de sets de discontinuidades

La caracterización de macizos rocosos para aplicaciones de ingeniería generalmente implica la identificación de conjuntos de discontinuidades y la caracterización de su orientación. Esto se hace comúnmente a través de redes estereográficas. Tradicionalmente, el análisis visual de gráficos de contorno de densidad, calculada contando el número de vectores normales unitarios que caen dentro de un círculo de referencia, se ha empleado para esta tarea. Sin embargo, se ha encontrado que este presentaba problemas debido al sesgo de muestreo, a la dependencia de los

resultados del tamaño del círculo de referencia y a la subjetividad en la interpretación de los resultados [10].

Estos problemas han llevado al desarrollo de técnicas alternativas para la identificación automática de sets de discontinuidades en función de su orientación. Por ejemplo, se han empleado métodos de *spectral clustering* que utilizan los vectores propios de matrices construidas mediante medidas de similitud entre los puntos de datos para el agrupamiento de las discontinuidades.

4.3 Nuevo método de *clustering* iterativo para sets de discontinuidades de rocas

El agrupamiento de discontinuidades es un paso importante y preliminar en la estimación de las propiedades de las discontinuidades. Para agrupar las discontinuidades de manera rápida y precisa, se propuso un nuevo método de agrupamiento iterativo (NICM) basado en los datos de las orientaciones de las discontinuidades y las longitudes de traza [11].

El método asume que las longitudes de traza de las discontinuidades siguen una distribución específica, a través de una redistribución de las discontinuidades cuyos valores de grado de pertenencia no están claros para ajustar la función de densidad de probabilidad de las longitudes de traza y hacer que se ajusten más estrechamente a la distribución específica.

Las características del método se analizaron con tres tipos de modelos numéricos en los que los datos son sintéticos y los resultados de agrupamiento se pueden comprender claramente.

Esta técnica será usada en este trabajo, al considerar un caso sintético para poner a prueba los algoritmos, para luego aplicarlos a un caso de estudio real.

4.4 Resumen

En resumen, hemos visto que la aplicación de algoritmos de clustering en la determinación de sets estructurales presenta desafíos significativos. En la literatura, se han identificado problemas relacionados con el sesgo de muestreo, la agrupación de resultados según el tamaño de referencia y la subjetividad en la interpretación de los resultados. Para abordar estos problemas, se han desarrollado y aplicado diversas técnicas, incluidos algoritmos de clustering difuso y métodos que utilizan distribuciones de datos específicas para ajustar la probabilidad de pertenencia a los clústeres. Estas técnicas buscan mejorar la precisión y la objetividad en la identificación de sets estructurales y es en esta línea del conocimiento donde se desarrolla este trabajo de memoria.

5 METODOLOGÍA

En el presente capítulo se presenta la metodología seguida para cumplir con los objetivos presentados en el capítulo 2.

5.1 Recopilación de Algoritmos

La primera etapa consiste en la selección de algoritmos. Esta selección se basa en algoritmos utilizados en la revisión bibliográfica, así como, algoritmos aplicados en la industria a través de softwares, tales como DIPS que utiliza *K-Means* [19].

También se incluyen algoritmos que no se encontró en la literatura que hayan sido utilizados con el fin de identificar sets estructurales.

Los algoritmos seleccionados corresponden a los siguientes:

1. *K-Means*
2. *Affinity Propagation*
3. *Spectral Clustering*
4. *Agglomerative Clustering*
5. *DBSCAN*
6. *Mean-Shift*

5.2 Implementación de Algoritmos

Los algoritmos seleccionados que se implementarán serán obtenidos *Scikit-learn* la cual es una librería de *machine learning* desarrollada en lenguaje Python. Estos algoritmos se encuentran optimizados.

Cada algoritmo de *clustering* se encuentra en dos variantes. La variante utilizada en este trabajo corresponde a una función que, dado un set de datos, devuelve una matriz de etiquetas (o *labels*) enteras correspondientes a los diferentes grupos (o *clusters*). Las etiquetas sobre los datos se pueden encontrar en el atributo *"labels_"* [21].

Las rutinas implementadas serán desarrolladas en Python, para la ejecución de los algoritmos mencionados implementados en *Scikit-learn.*

5.3 Procesamiento de Datos

Debido a que los algoritmos tienen su métrica definida al estar programados en las librerías de *scikit-learn*, se decide utilizar un pre-procesamiento de los datos con tal de emular una métrica angular, al momento que el algoritmo calcule la distancia euclidiana. Este pre-procesamiento consiste en la transformación de los datos de dip y dipdir a otro espacio, ya sea R^2, R^3 o superior.

Las transformaciones utilizadas para el pre-procesamiento, se presentan a continuación:

5.3.1 Transformación trigonométrica en R^2

Esta transformación fue propuesta debido a que $dipdir$ de 0° y de 359° deberían poseer una distancia pequeña, pues son datos angulares. Esto ocurre debido a que $sin(0°)$ es muy cercano a $sin(350°)$. El tomar $sin(dip)$ se justifica para que los valores estuvieran en el mismo orden de magnitud.

La transformación es la siguiente:

$$\begin{pmatrix} Dip \\ DipDir \end{pmatrix} - -> \begin{cases} sin^2(DipDir) \\ \sin(Dip) \end{cases}$$

5.3.2 Transformación esférica a R^3

Esta transformación fue propuesta por Hammah y Curran en el 2000. Está basada en una transformación esférica de los datos, pues estos se representan en una semi esfera [9].

La transformación es la siguiente:

$$\begin{pmatrix} Dip \\ DipDir \end{pmatrix} - -> \begin{cases} \sin(DipDir)\sin(Dip) \\ \cos(DipDir)\sin(Dip) \\ \cos(Dip) \end{cases}$$

5.3.3 Transformación trigonométrica a R^4

Esta transformación sigue el mismo principio anterior, para compensar la distancia euclidiana con los datos angulares, pero incluye la noción de que al aumentar la dimensionalidad del problema (en este de R^2 a R^4), se obtiene un espacio en el cual existe un hiperplano que divide a conjunto de datos, lo cual favorece el *clustering*.

La transformación es la siguiente:

$$\begin{pmatrix} Dip \\ DipDir \end{pmatrix} - -> \begin{cases} \sin(Dip) \\ \cos(Dip) \\ \sin(DipDir) \\ \cos(DipDir) \end{cases}$$

5.4 Aplicación de Algoritmos

Se aplicarán los siguientes algoritmos: *K-Means*, *Affinity Propagation*, *Spectral Clustering*, *Agglomerative Clustering*, *DBSCAN* y *Mean-Shift*, con las rutinas programadas previamente.

Cada uno de estos algoritmos será aplicado a los datos preprocesados, y las etiquetas resultantes indicarán a qué clúster pertenece cada discontinuidad. Esta aplicación permitirá comparar cómo los diferentes algoritmos de clustering agrupan las discontinuidades y si estos clústers son coherentes con los esperados.

5.5 Visualización de Resultados

Después de aplicar los algoritmos de clustering a los datos de discontinuidades, los resultados se representarán en una red estereográfica. En esta representación, cada punto corresponderá a una orientación de discontinuidad, y se utilizarán diferentes colores para indicar a qué cluster pertenece dicha discontinuidad. Los colores permitirán una identificación clara de los clusters, lo que facilitará la interpretación de los resultados.

5.6 Evaluación del Desempeño de Algoritmos

La evaluación del desempeño de los algoritmos se realizará de forma visual para un conjunto de datos sintéticos, cuyos sets son conocidos [23].

La visualización de los resultados se realizará utilizando la librería *mplstereonet* para las redes estereográficas y los diagramas de contorno. *mplstereonet* es una librería proporciona redes estereográficas de igual área y ángulo de hemisferio inferior para *matplotlib*[1]. También proporciona algunos métodos diferentes para producir diagramas de contorno [8].

Los diagramas de contorno son realizados mediante el método de Kamb modificado [25].

Adicionalmente, se evaluará el número adecuado u óptimo de clústeres mediante de técnicas como el método del codo y el método de la silueta.

También se evaluará el desempeño a través de métricas como el índice de silueta y el índice Calinski-Harabasz.

5.7 Casos de Estudio

Los algoritmos que presenten un mejor desempeño para el caso sintético serán aplicados a una base de datos real, que incorpora las variables de dip y dipdir georeferenciadas. Esto permite visualizar no sólo los sets estructurales en las redes estereográficas, sino también su distribución y continuidad espacial.

Finalmente, en el siguiente esquema se representa un resumen de la metodología expuesta.

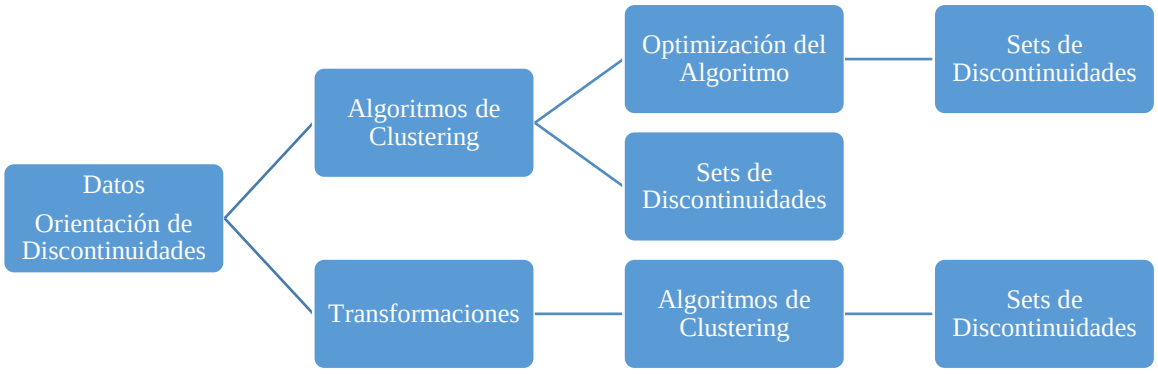

Figura 5-1 Esquema de la metodología utilizada.

[1] *Matplotlib* es una biblioteca completa para crear visualizaciones estáticas, animadas e interactivas en Python

6 RESULTADOS CASO SINTÉTICO

La primera aproximación para evaluar el desempeño de los algoritmos de *clustering* será trabajar con un caso sintético.

6.1 Datos

Los datos utilizados provienen de una base de 144 datos de orientación de discontinuidades caracterizados por su Dip y DipDirection. A continuación, se presentan los datos graficados en una red estereográfica. La proyección se realiza en el hemisferio inferior, y con ángulos iguales.

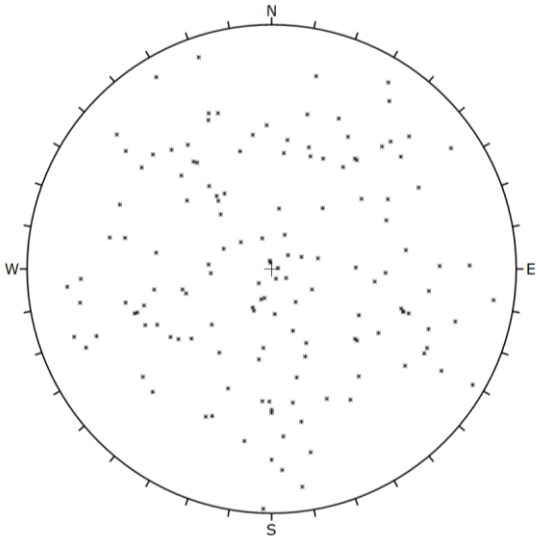

Figura 6-1 Representación polos Caso Sintético. Realizado en software DIPS.

A continuación, se presentan los diagramas de contorno, para ver los indicios de los potenciales sets de discontinuidades. Estos diagramas fueron realizados con el software Dips y con la librería *mplstereonet* para Python. Ambos resultados son consistentes.

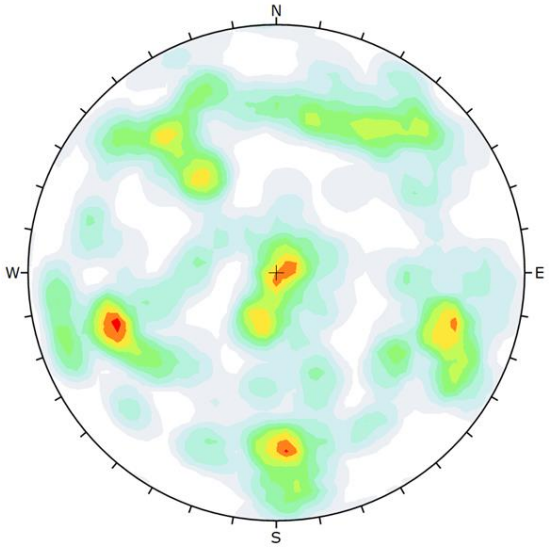

Figura 6-2 Diagrama de Contorno. Caso Sintético. Realizado en software DIPS.

Color	Density Concentrations
	0.00 - 0.40
	0.40 - 0.80
	0.80 - 1.20
	1.20 - 1.60
	1.60 - 2.00
	2.00 - 2.40
	2.40 - 2.80
	2.80 - 3.20
	3.20 - 3.60
	3.60 - 4.00

Contour Data	Pole Vectors
Maximum Density	3.85%
Contour Distribution	Fisher
Counting Circle Size	1.0%

Plot Mode	Pole Vectors
Vector Count	144 (144 Entries)
Hemisphere	Lower
Projection	Equal Area

Figura 6-3 Leyenda Diagrama de Contorno realizado en software DIPS.

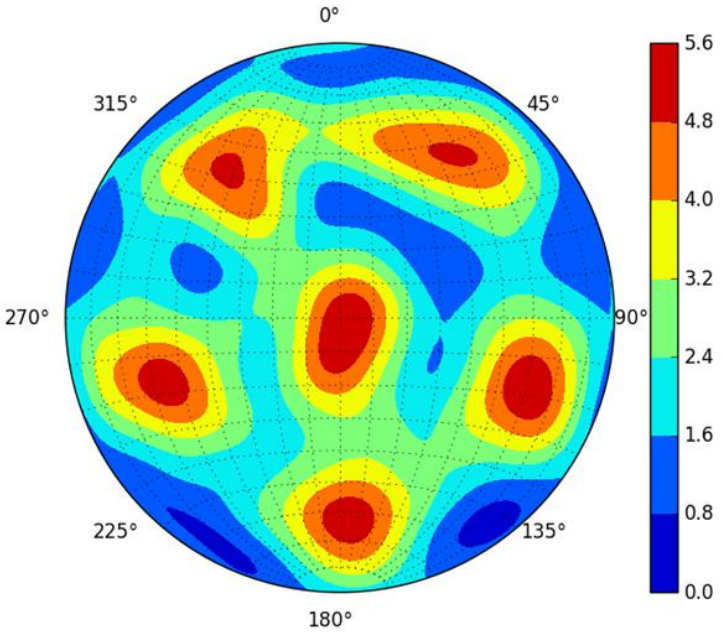

Figura 6-4 Diagrama de Contorno. Caso Sintético. Realizado con la librería *mplstereonet* en Python.

En la Figura 6-4 se observa que los clústers esperados corresponden a 6. Un clúster central y 5 clústers en el borde.

A continuación, se presenta una representación en 3D del diagrama de contorno y polos.

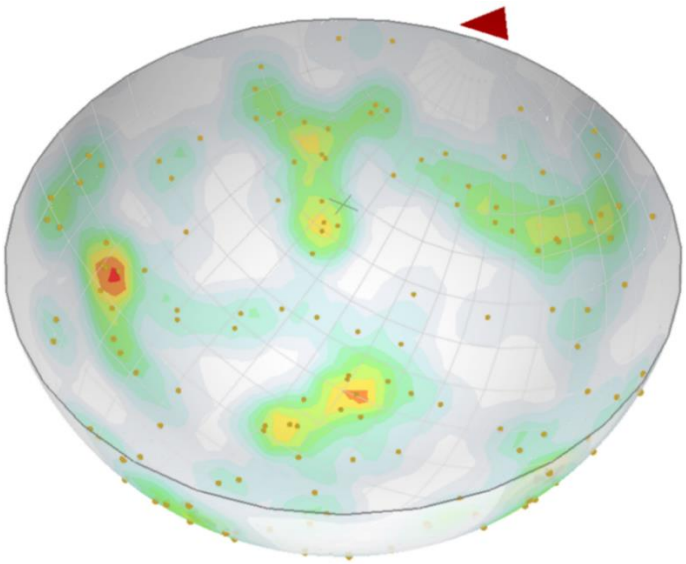

Figura 6-5 Representación 3D. Proyección en la esfera inferior.

6.2 Resultados sin transformaciones

A continuación, se presentan los resultados más importantes, para los algoritmos, sin transformaciones.

6.2.1 K-Means

Los resultados que más se condicen con los esperados corresponden a 5 y 6 clústers.

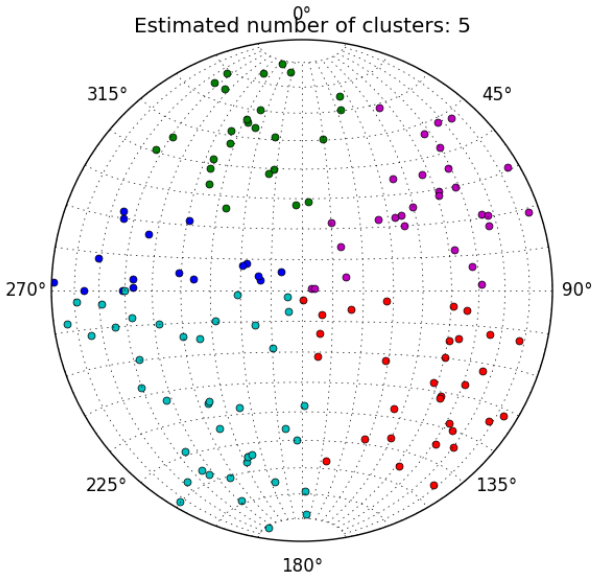

Figura 6-6 Representación Gráfica Polos Resultado *K-Means* sin transformación. Número de Clústeres = 5.

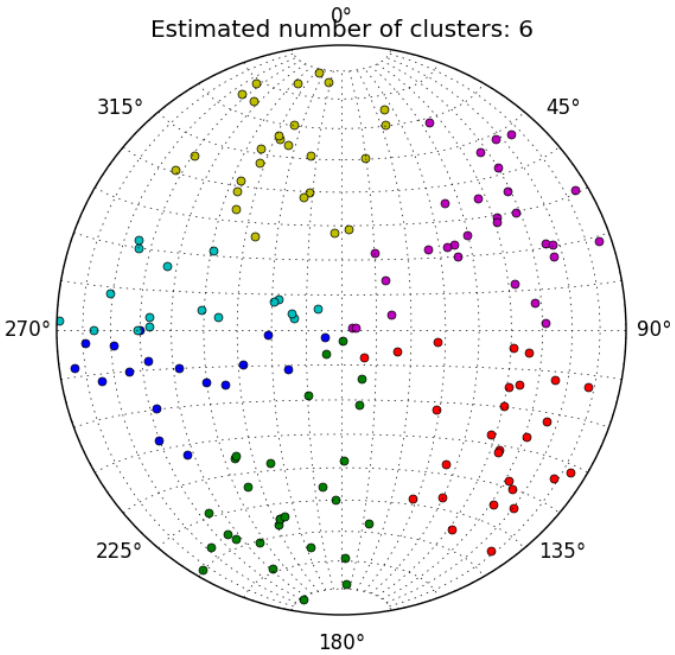

Figura 6-7 Representación Gráfica Polos Resultado *K-Means* sin transformación. Número de Clústeres = 6.

6.2.2 DBSCAN

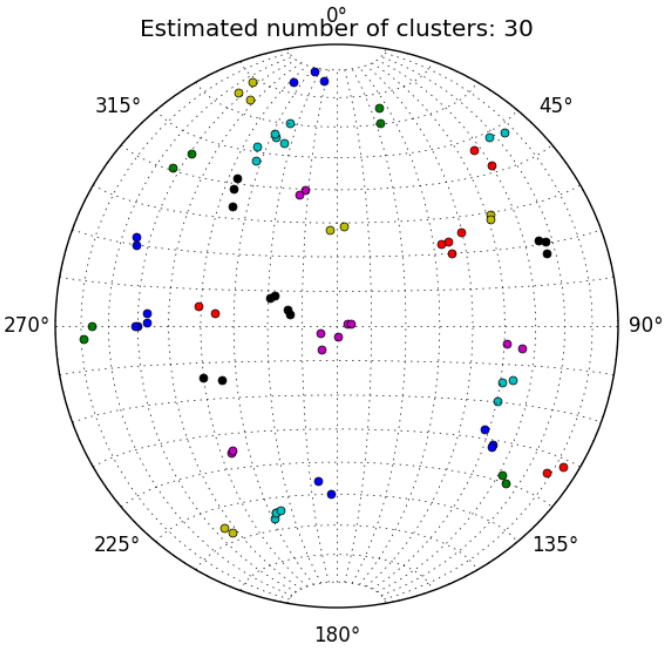

Figura 6-8 Representación Gráfica Polos Resultado DBSCAN sin transformación. Número de Clústeres = 30. Los polos que no están asociado a un clúster no se presentan en la red estereográfica.

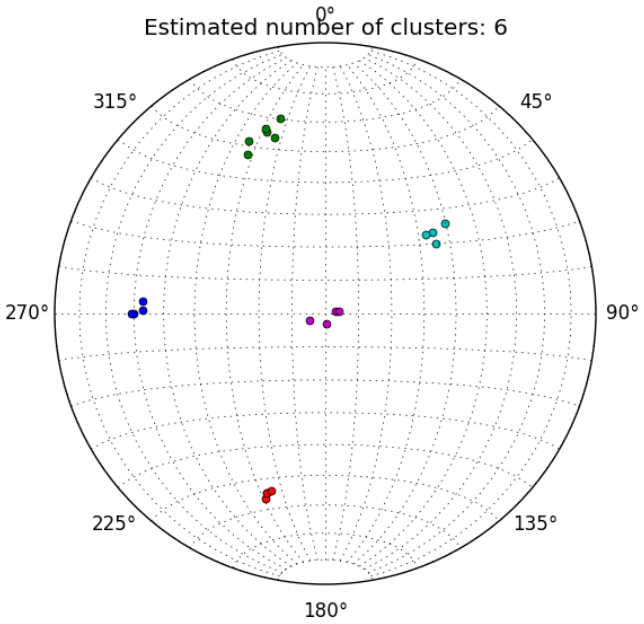

Figura 6-9 Representación Gráfica Polos Resultado DBSCAN sin transformación. Número de Clústeres = 6. Los polos que no están asociado a un clúster no se presentan en la red estereográfica.

6.2.3 Agglomerative Clustering

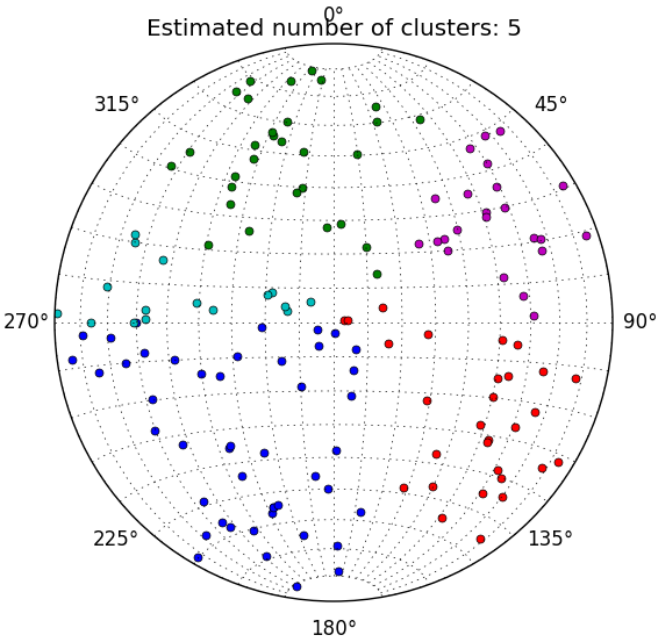

Figura 6-10 Representación Gráfica Polos Resultado *Agglomerative Clustering* sin transformación. Número de Clústeres = 5.

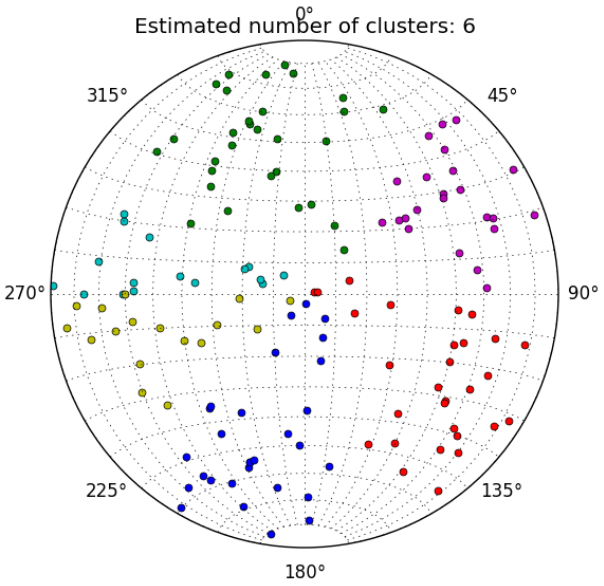

Figura 6-11 Representación Gráfica Polos Resultado Agglomerative Clustering sin transformación. Número de Clústeres = 6.

Para ningún algoritmo se obtuvo el resultado esperado, de acuerdo con el diagrama de contorno, que muestra que son 6 sets, y uno de ellos está en el centro de la red estereográfica.

Estos resultados indican la necesidad de considerar transformaciones a los datos para mejorar el rendimiento de los algoritmos y la precisión en la identificación de los sets estructurales.

A continuación, se presentarán los resultados después de aplicar las transformaciones a los datos de orientación de discontinuidades.

6.3 Resultados con transformación trigonométrica a R^2 y R^4.

Los resultados con las transformaciones trigonométricas a R^2 y R^4 no logran identificar los 6 sets de discontinuidades esperados. A pesar de aplicar transformaciones a los datos, los algoritmos de clustering aún muestran un rendimiento deficiente en la identificación de los sets estructurales.

Estos resultados indican que las transformaciones trigonométricas a R^2 y R^4 no son efectivas para mejorar la precisión del clustering en este caso.

6.4 Resultados con transformación esférica a R^3

Los resultados anteriores indican la necesidad de utilizar otro tipo de transformación. A continuación, se presentarán los resultados después de aplicar la transformación esférica a R^3 propuesta por Hammah y Curran en el 2000.

A continuación, se presenta una visualización 3D en el sistema cartesiano donde se puede ver por qué se denomina transformación esférica. A cada una de las transformaciones se le denominará Variable 1, Variable 2 y Variable 3.

$$\begin{pmatrix} Dip \\ DipDir \end{pmatrix} --> \begin{cases} \sin(DipDir)\sin(Dip) \rightarrow (Variable1) \\ \cos(DipDir)\sin(Dip) \rightarrow (Variable2) \\ \cos(Dip) \rightarrow (Variable3) \end{cases}$$

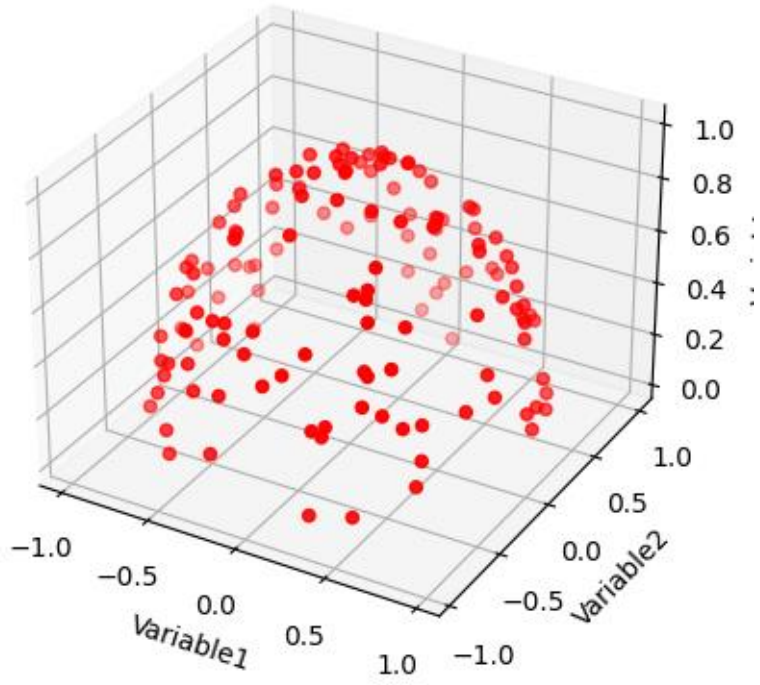

Figura 6-12: Visualización 3D. Datos aplicando la transformación esférica propuesta por Hammah y Curran en el 2000 (Elaboración propia).

6.4.1 K-Means

En este caso, donde el único parámetro de entrada es el número de clústeres, se reconocen los 5 clústeres laterales, más el clúster central esperado.

Las estadísticas promedio por cluster se presentan en la Tabla 6-1. La representación en la red estereográfica se presenta en la Figura 6-13, mientras que la representación 3D de los datos transformados se presenta en la Figura 6-14.

Tabla 6-1: Estadísticas de DipDir y Dip para cada cluster.

Cluster	DipDir Promedio [°]	Dip Promedio [°]
0	182.79	16.67
1	289.38	62.54
2	178.67	60.92
3	215.25	63.08
4	71.46	61.04
5	145.38	60.42

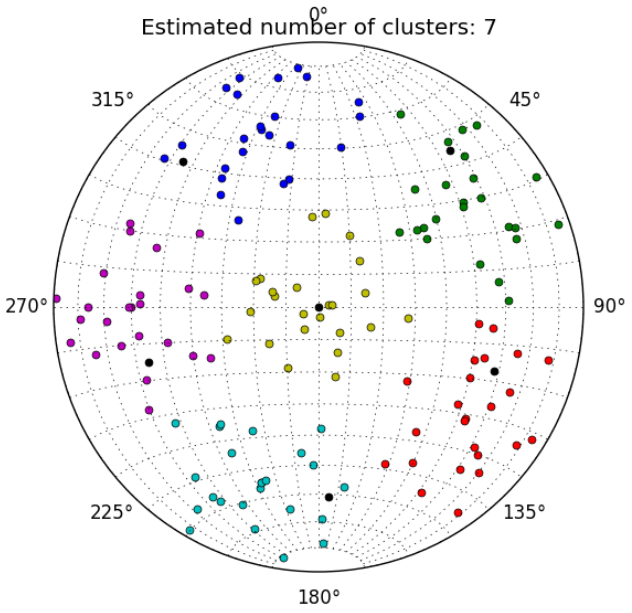

Figura 6-13 Representación Gráfica Polos Resultado *K-Means* con transformación esférica. Número de Clústeres = 6.

Otra forma de ver los clusters es graficarlos en el sistema de ejes cartesianos.

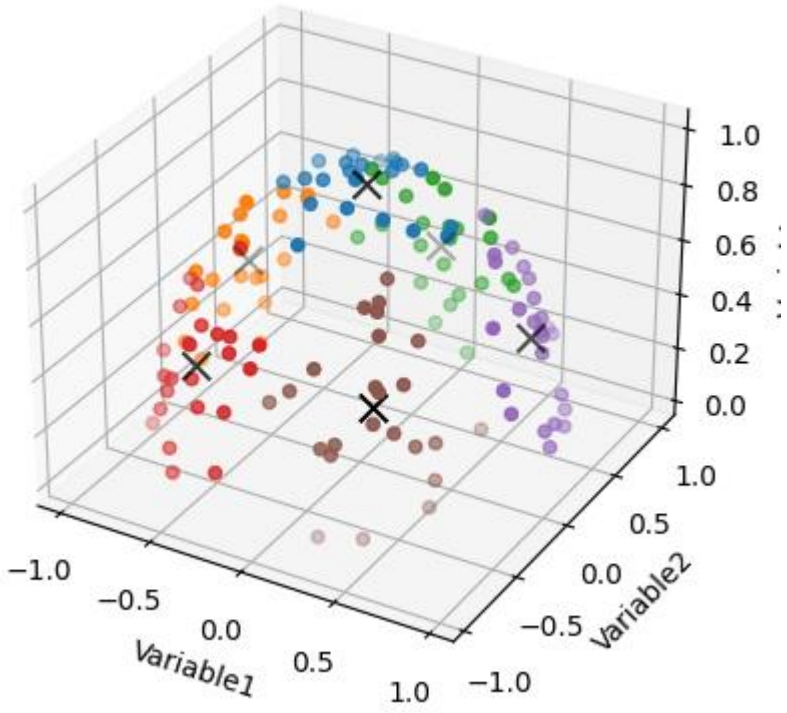

Figura 6-14: Representación Gráfica 3D Resultado K-Means con datos transformados. Número de Clústeres = 6 (Elaboración propia).

6.4.2 Spectral Clustering

De igual forma que con *K-Means*, para el *spectral clustering* el único parámetro de entrada es el número de clústeres, y se obtiene el mismo resultado que en el caso anterior, que es el resultado esperado.

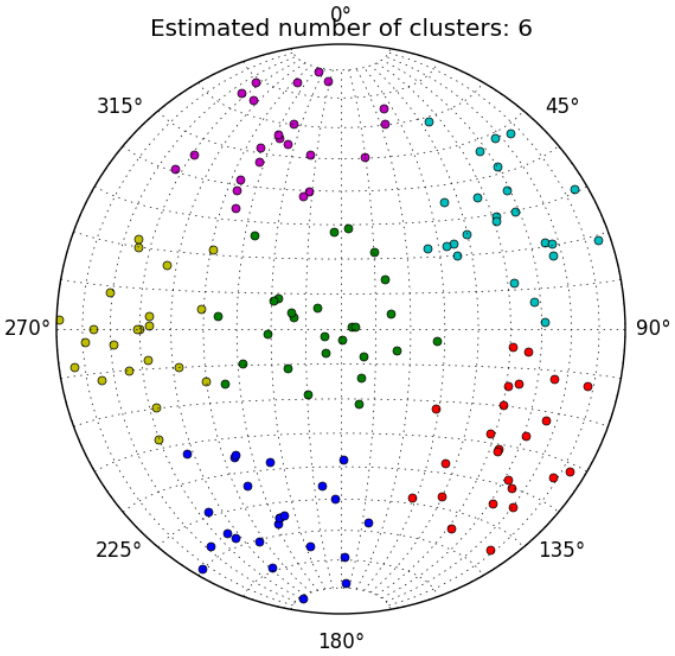

Figura 6-15 Representación Gráfica Polos Resultado *Spectral Clustering* con transformación esférica. Número de Clústeres = 6.

6.4.3 Affinity Propagation

En este caso, para el *affinity propagation* el parámetro de entrada es el *damping*, que se puede interpretar como un radio de búsqueda, se obtiene una sobre-partición de los datos, donde se obtienen 7 clústeres. Cabe destacar que se reconoce el clúster central, que es el principal que presenta problemas para su reconocimiento.

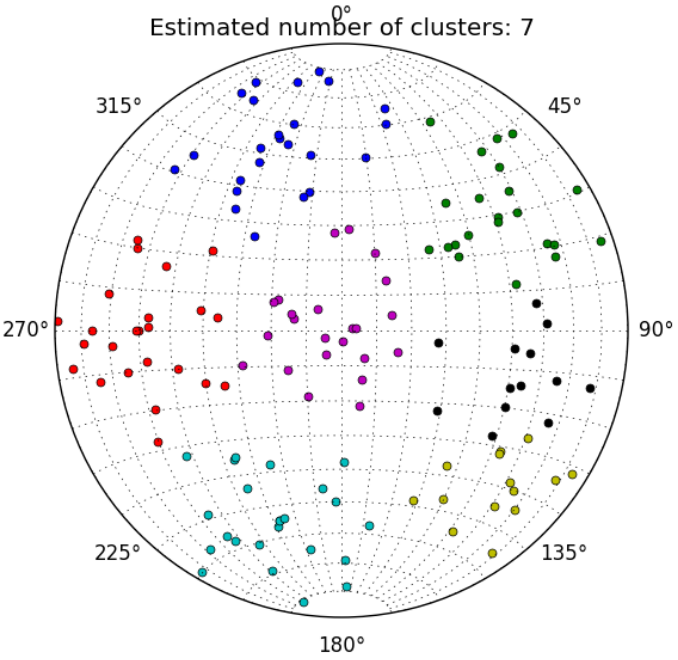

Figura 6-16 Representación Gráfica Polos Resultado *Affinity Propagation* con transformación esférica. Número de Clústeres = 6.

En resumen, para estos casos se observa que, los algoritmos cuyo único parámetro el número de clústeres, entregan un resultado acorde al diagrama de contorno que indica que son 6 clústeres, y uno de ellos está al centro de la red estereográfica.

6.5 Evaluación del número óptimo de clústers

La primera técnica que se revisará será el método del codo.

El gráfico para el algoritmo K-Means muestra la relación entre la cantidad de clústeres en el eje horizontal y la dispersión intraclúster en el eje vertical. La dispersión intraclúster se refiere a la variabilidad de los puntos de datos dentro de cada clúster.

Se espera que al aumentar el número de clusters, la dispersión intraclúster disminuya, ya que los datos se dividen en un mayor número de grupos más pequeños. Sin embargo, se llegará a un punto en el que el aumento adicional del número de clusters no conlleve una mejora significativa, y la dispersión intraclúster se estabilizará. Es decir, se debe buscar el punto de quiebre en la pendiente o el codo.

El gráfico con el método del codo se presenta en la Figura 6-17. En dicha figura se observa que el punto de quiebre se produce en los 6 clústers, que se condice con lo esperado para el caso sintético.

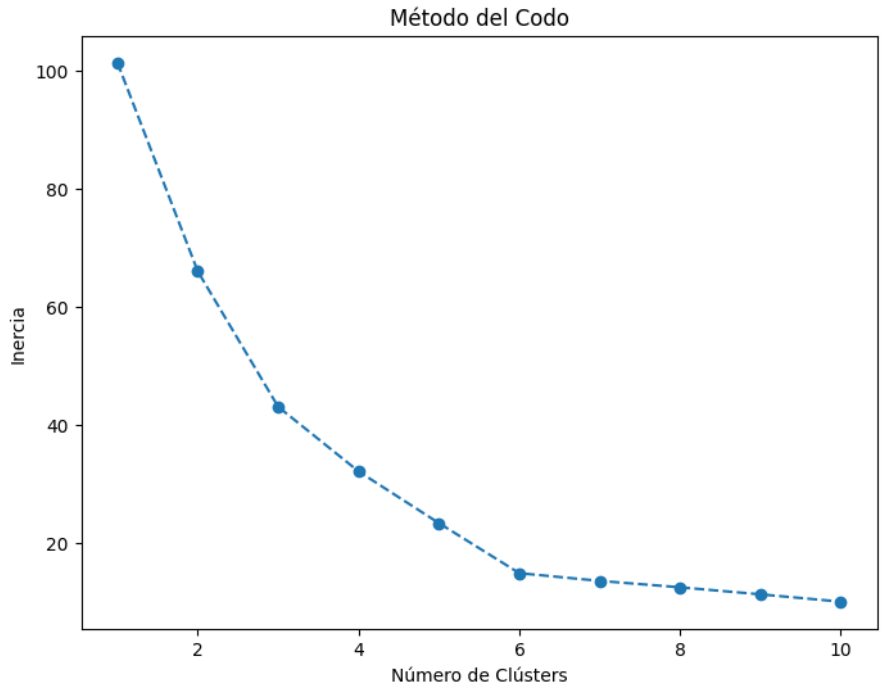

Figura 6-17: Método del codo aplicado con el K-Means

La segunda técnica que se revisará será el método de la silueta.

El gráfico con el método de la silueta se presenta en la Figura 6-18. En dicha figura se observa que el máximo se produce en los 6 clústers, que se condice con lo esperado para el caso sintético.

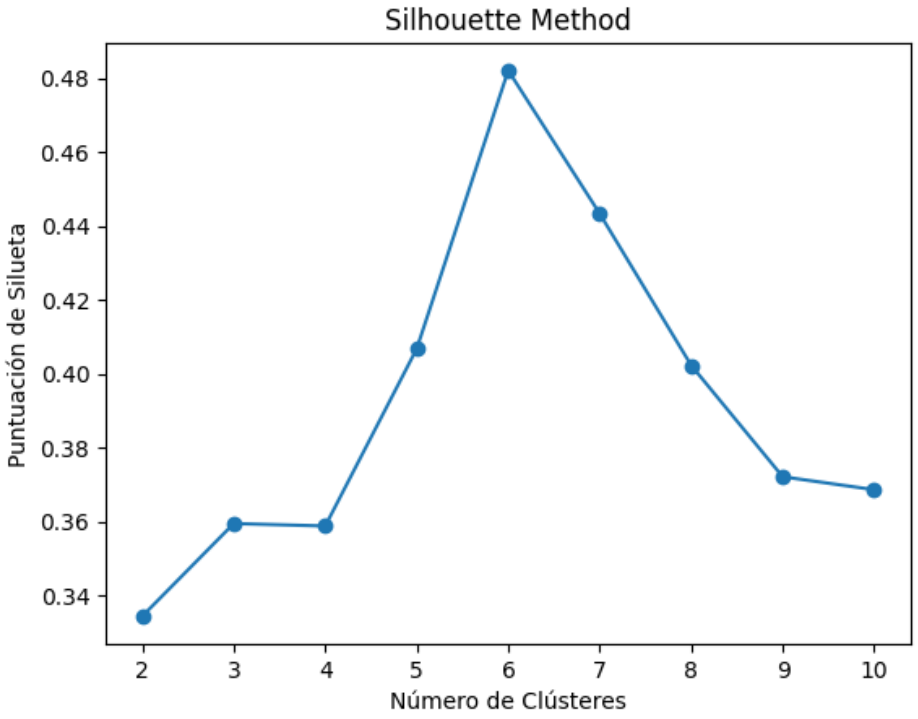

Figura 6-18: Método de la silueta aplicado con el K-Means.

La tercera técnica que se revisará será el método Gap Statistic.

El gráfico con el método Gap Statistic se presenta en la Figura 6-19. En dicha figura se observa que la máxima diferencia entre el *gap value* se produce al pasar de los 5 a los 6 clústers, por lo que el número óptimo de clústers corresponde a 6, que se condice con lo esperado para el caso sintético.

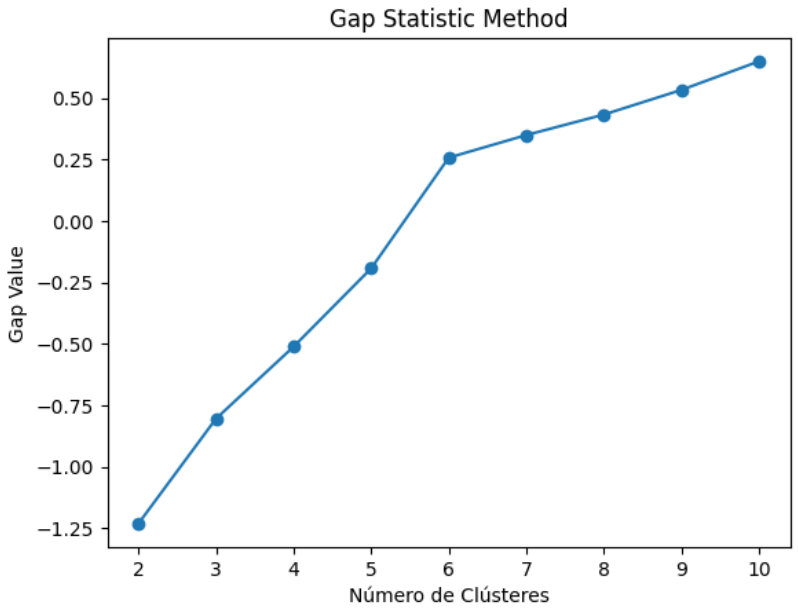

Figura 6-19: Método Gap Statistic aplicado con el K-Means.

6.6 Métricas del desempeño

Las métricas de desempeño para el algoritmo *K-Means* aplicado a los datos transformados en R^3 (mediante la transformación esférica) con 6 clústers se presentan en la Tabla 6-2, y son las siguientes: el índice de silueta, el índice Índice Calinski-Harabasz y la puntuación de Dunn.

Tabla 6-2: Métricas de desempeño para K-means aplicado al caso sintético.

Métrica	**Valor**
Índice de Silueta (Silhouette Score)	0.48
Índice Calinski-Harabasz	160.31
Puntuación de Dunn	0.14

7 RESULTADOS CASOS DE ESTUDIO

Los resultados de los capítulos anteriores nos permiten seleccionar al algoritmo *K-Means* como el que presenta mejor desempeño. Dicho algoritmo, incorporando la selección de centroides para su inicialización, será el utilizado para estudiar 2 casos de estudio reales. Las bases de datos se encuentran georreferenciadas, lo que permitirá evaluar el desempeño de los algoritmos considerando la variable espacial.

7.1 Caso de Estudio 1

La base de datos consta de 47.584 datos. Los datos provienen de la pared de un rajo. En la Tabla 7-1 se presenta las estadísticas básicas para el Dip. Se observa que no existen valores aberrantes pues todos los datos se encuentran en el intérvalo de 0 a 90°. La distribución acumulada se presenta en el histograma de la Figura 7-1.

Tabla 7-1 Estadísticas básicas Dip. Caso de Estudio 1.

Estadístico	Valor Dip [°]
Media	56,05
Desviación Estándar	22,53
Mínimo	0,52
Máximo	90,00

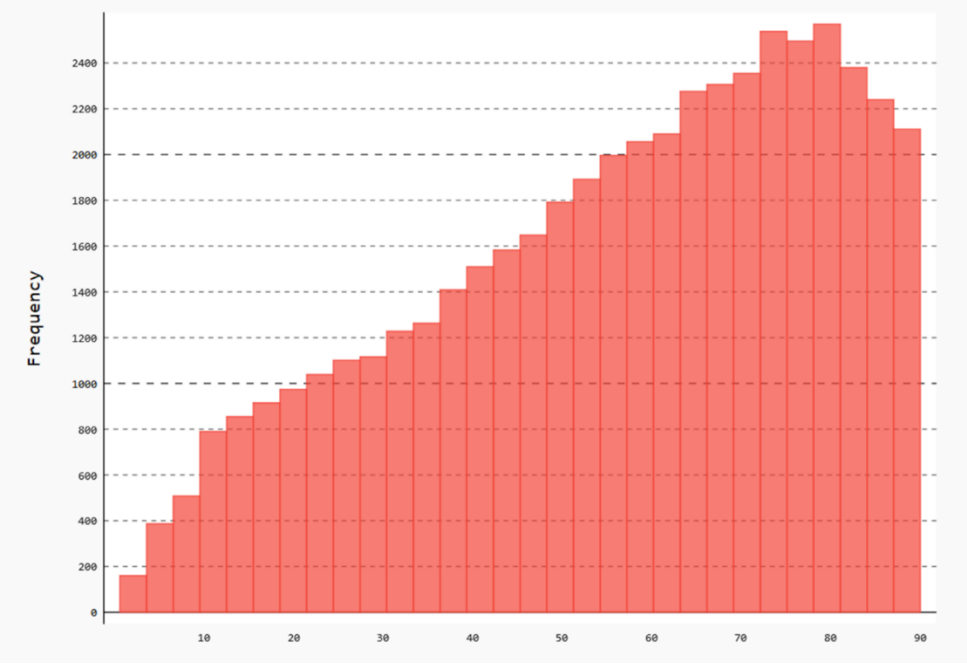

Figura 7-1 Histograma acumulado Dip. Caso de Estudio 1.

En la Figura 7-2 se presenta el diagrama de contorno para el Caso de Estudio 1. Se identifican 3 zonas de alta densidad, por lo que se espera que los resultados de los sets clasifiquen 3 grupos.

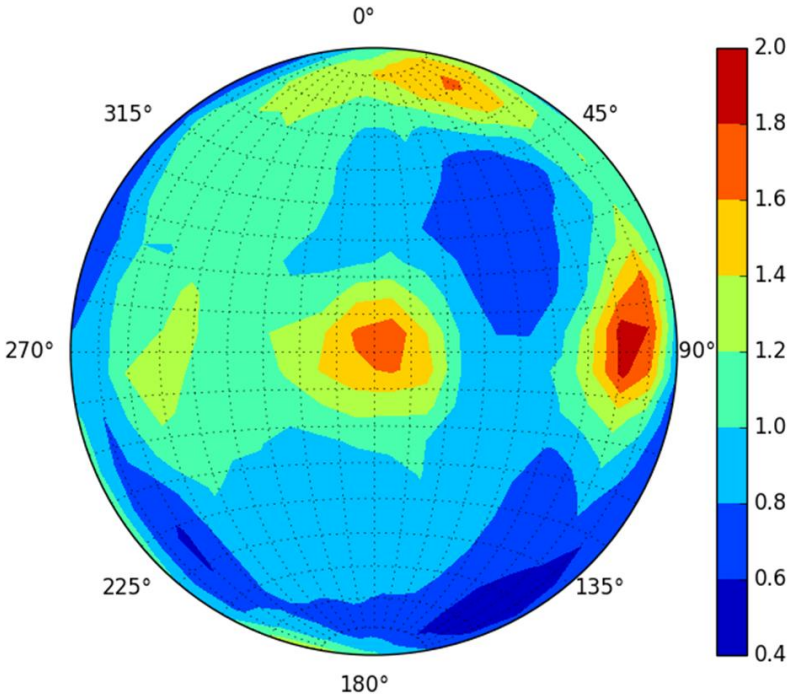

Figura 7-2 Diagrama de Contorno. Caso de Estudio 1.

Como se mencionó en el capítulo anterior, para la elección del número de centroides se requiere de los parámetros *Grid* y *Threshold*. Para los resultados presentados a continuación corresponden a una $Grid = 40$ y un $Threshold = 95$.

Tabla 7-2 Parámetros *Grid* y *Threshold* para selección de centroides.

Caso de Estudio 1	
Grid	*Threshold*
20	95 / 98
40	95 / 98
60	95 / 98

Los centroides para inicializar el algoritmo se presentan en la Figura 7-3 encerrados en color rojo. Se observa que los centroides coinciden con zonas de alta densidad de polos, como es requerido. Se debe mencionar también que a pesar de que los centroides no se hayan seleccionado exactamente en las zonas de alta densidad, esto se verá mitigado al ejecutar el algoritmo que moverá el centroide en cada iteración.

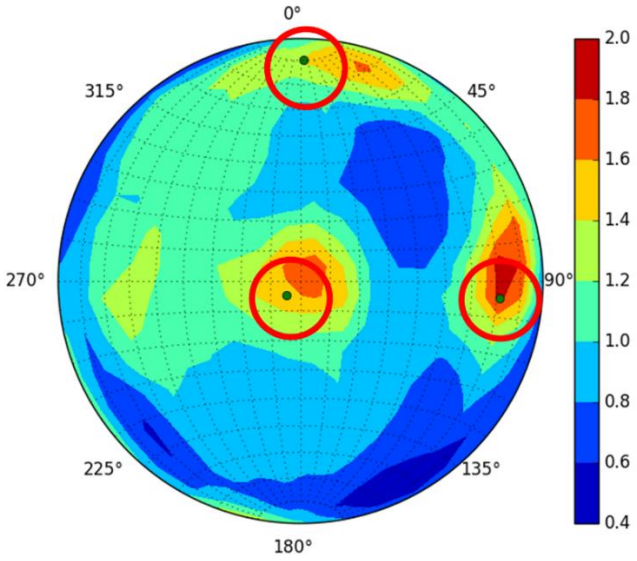

Figura 7-3 Número y ubicación de centroides para inicializar *K-Means*. Caso de Estudio 1.

Los resultados de los sets obtenidos se presentan a través de diagramas de contorno que sólo consideran los polos que perteneces a cada clúster en particular. Los diagramas se presentan en la Figura 7-5, Figura 7-6 y Figura 7-7, para el Clúster 1, Clúster 2, y Clúster 3, respectivamente.

En la Figura 7-4 se presenta el histograma con la cantidad de polos para cada 1 de los 3 *clusters*. La distribución tiende a una uniforme, donde no se tiene preferencia por ninguna orientación en particular.

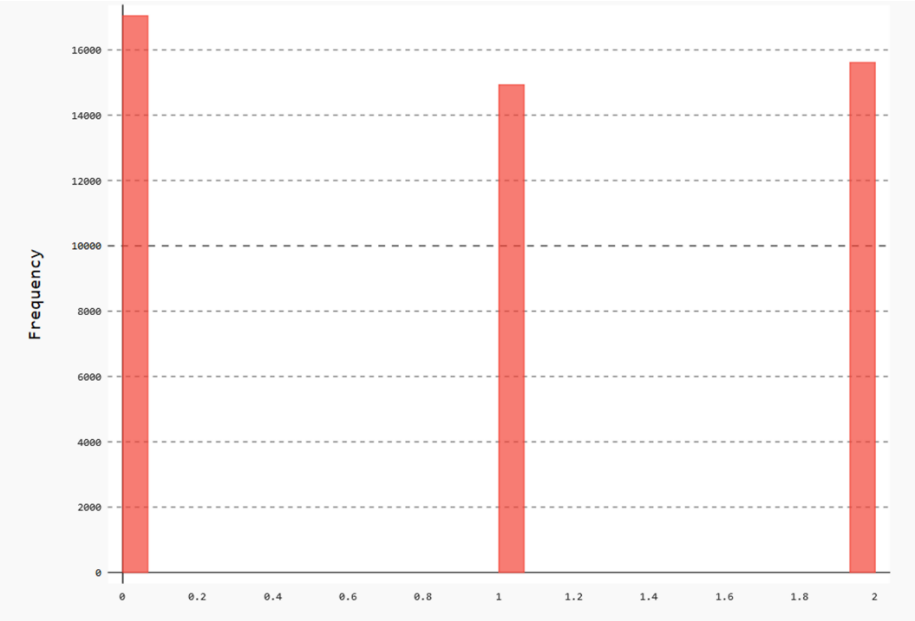

Figura 7-4 Histograma *Clusters*. Caso de Estudio 1.

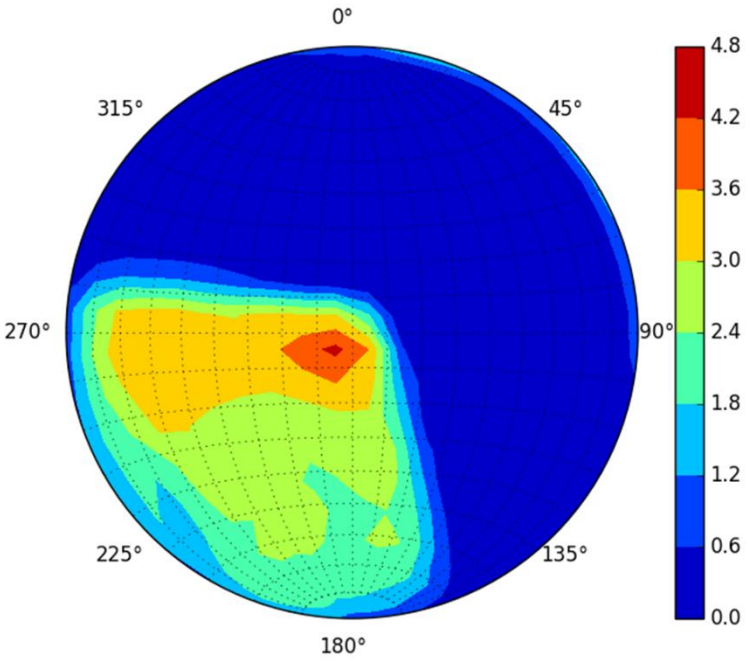

Figura 7-5 *Clúster* Número 1. Caso de Estudio 1.

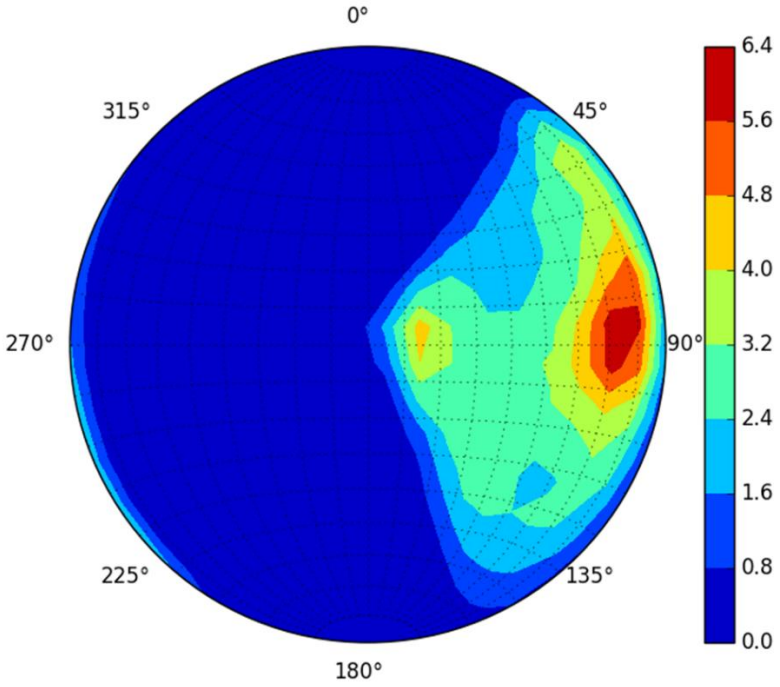

Figura 7-6 Clúster Número 2. Caso de Estudio 1.

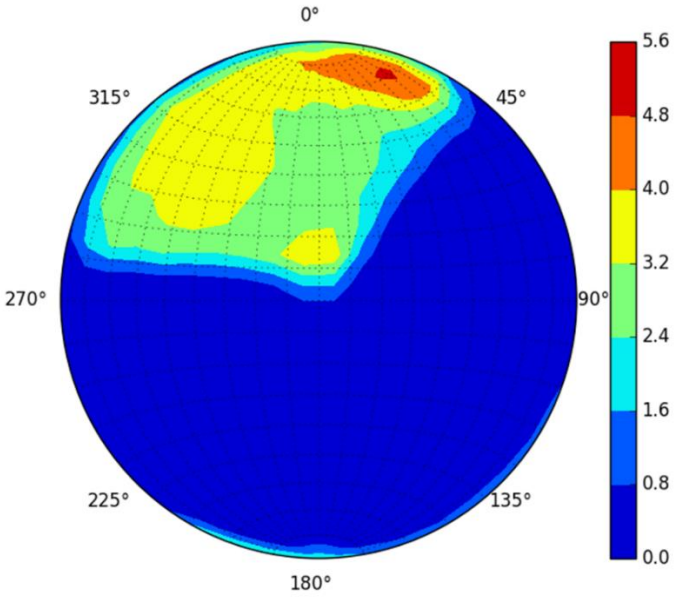

Figura 7-7 Clúster Número 3. Caso de Estudio 1.

Finalmente, se visualiza la distribución espacial de los sets. En la Figura 7-8 se observa una vista en planta de los datos. La visualización de las discontinuidades se realizó en el *software Leapfrog*. No se observa una correlación entre clúster y ubicación espacial de los datos.

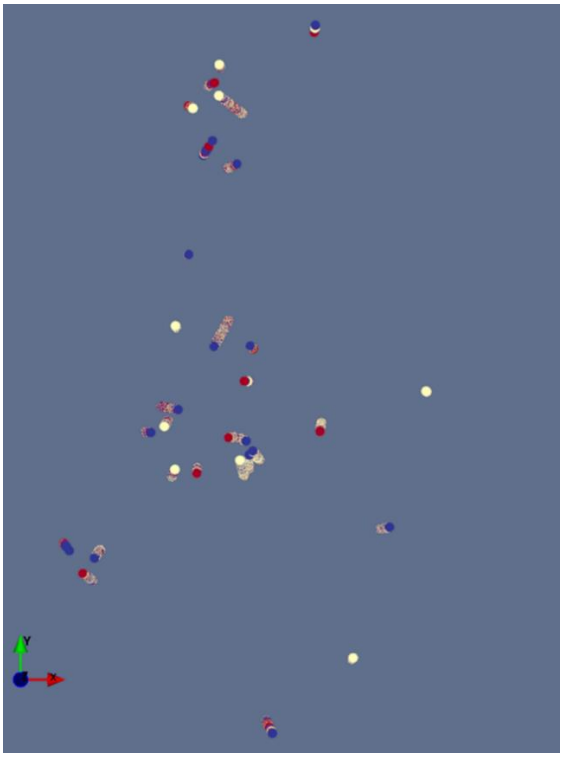

Figura 7-8 Distribución Espacial Clúster[2]. Vista en Planta. Caso de Estudio 1.

7.2 Caso de Estudio 2

La base de datos consta de 22.270 datos. En la Tabla 7-3 se presenta las estadísticas básicas para el Dip. Se observa que no existen valores aberrantes pues todos los datos se encuentran en el inérvalo de 0 a 90°. La distribución acumulada se presenta en el histograma de la Figura 7-9. A diferencia, del Caso de Estudio 1 se observa una distribución con una mayor cantidad de estructuras subverticales.

Tabla 7-3 Estadísticas básicas Dip. Caso de Estudio 2.

Estadístico	Valor Dip [°]
Media	63,19
Desviación Estándar	18,33
Mínimo	0,57
Máximo	90,00

[2] Cada color de punto representa uno de los 3 clústeres identificados.

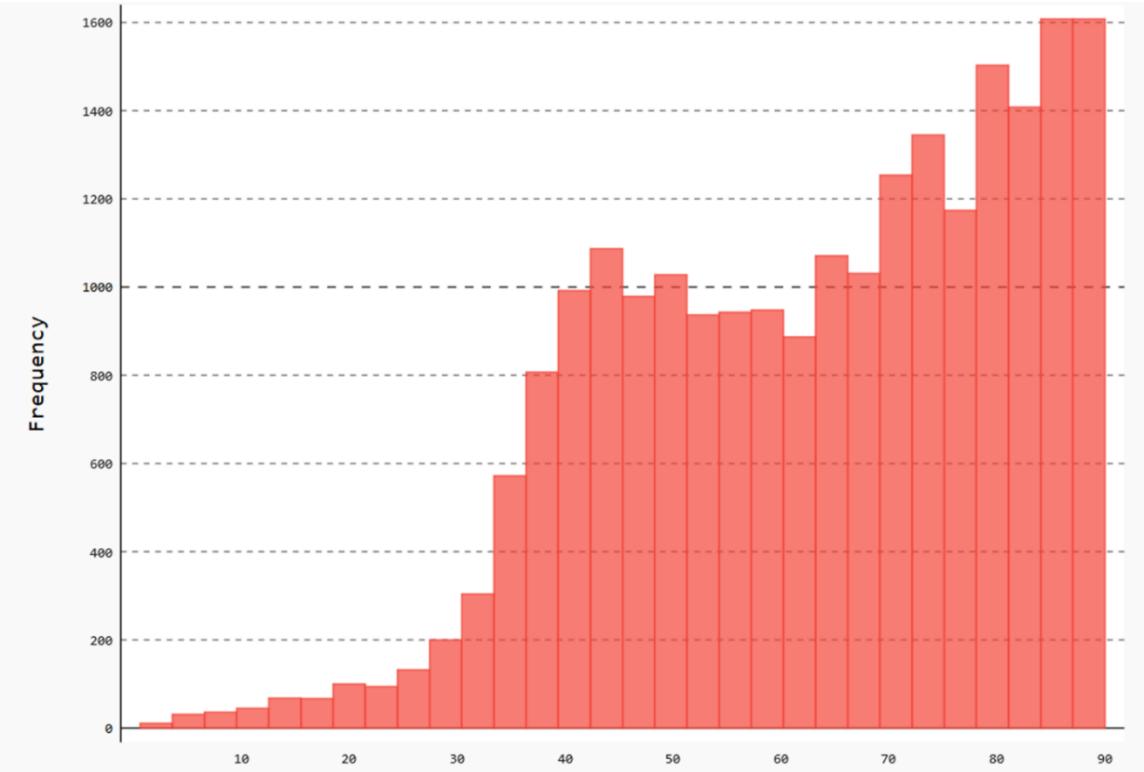

Figura 7-9 Histograma acumulado Dip. Caso de Estudio 2.

El diagrama de contorno para el Caso 2 se presenta en la Figura 7-10

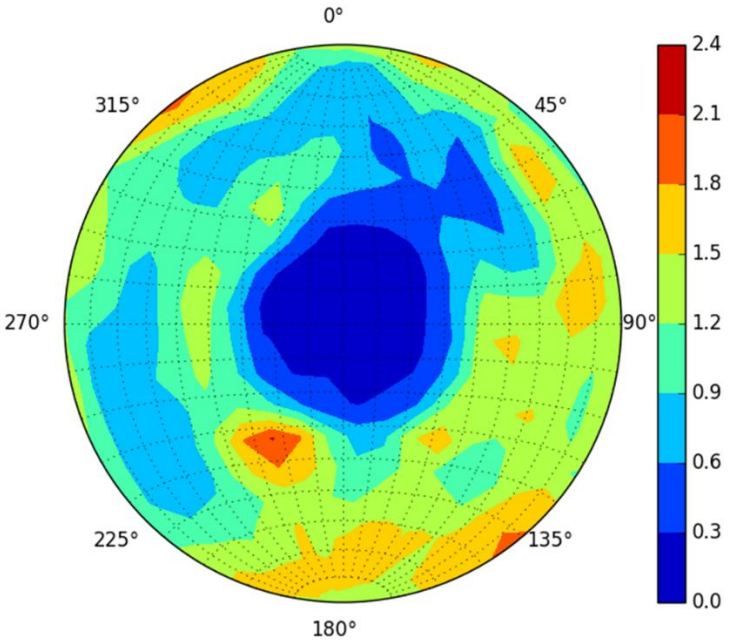

Figura 7-10 Diagrama de Contorno. Caso de Estudio 2.

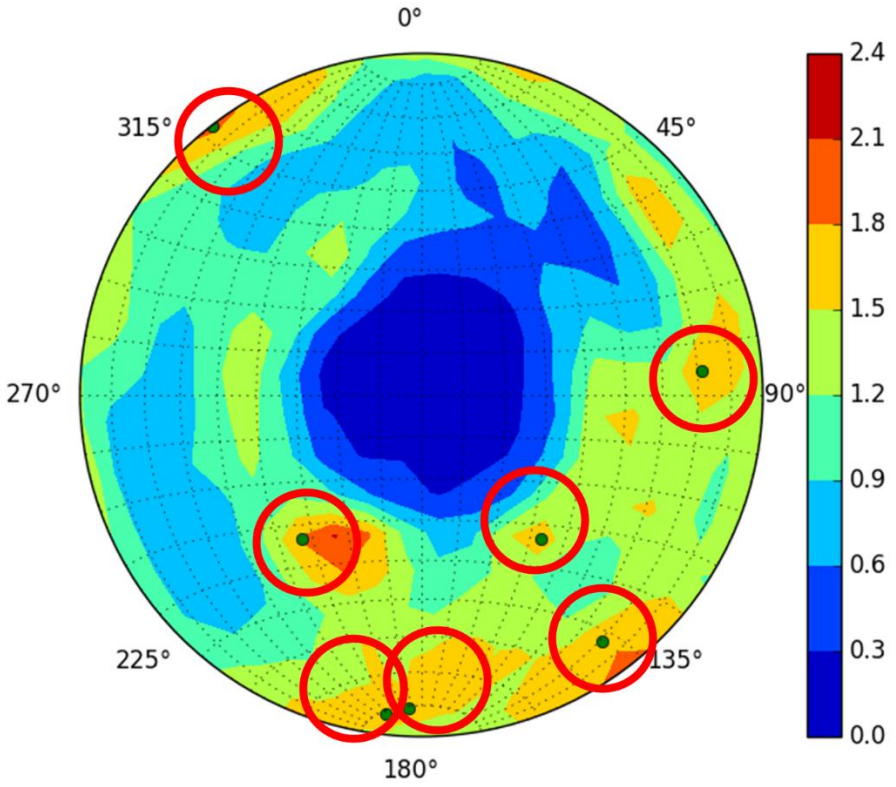

Figura 7-11 Número y ubicación de centroides para inicializar K-Means. Caso de Estudio 2.

Los resultados de los sets obtenidos se presentan a través de diagramas de contorno que sólo consideran los polos que perteneces a cada clúster en particular. Los diagramas se presentan desde la Figura 7-13 hasta la Figura 7-19, para los Clústeres desde el 1 hasta el 7.

En la Figura 7-12 se presenta el histograma con la cantidad de polos para cada 1 de los 7 *clusters*. La distribución tiende a una uniforme, donde no se tiene preferencia por ninguna orientación en particular.

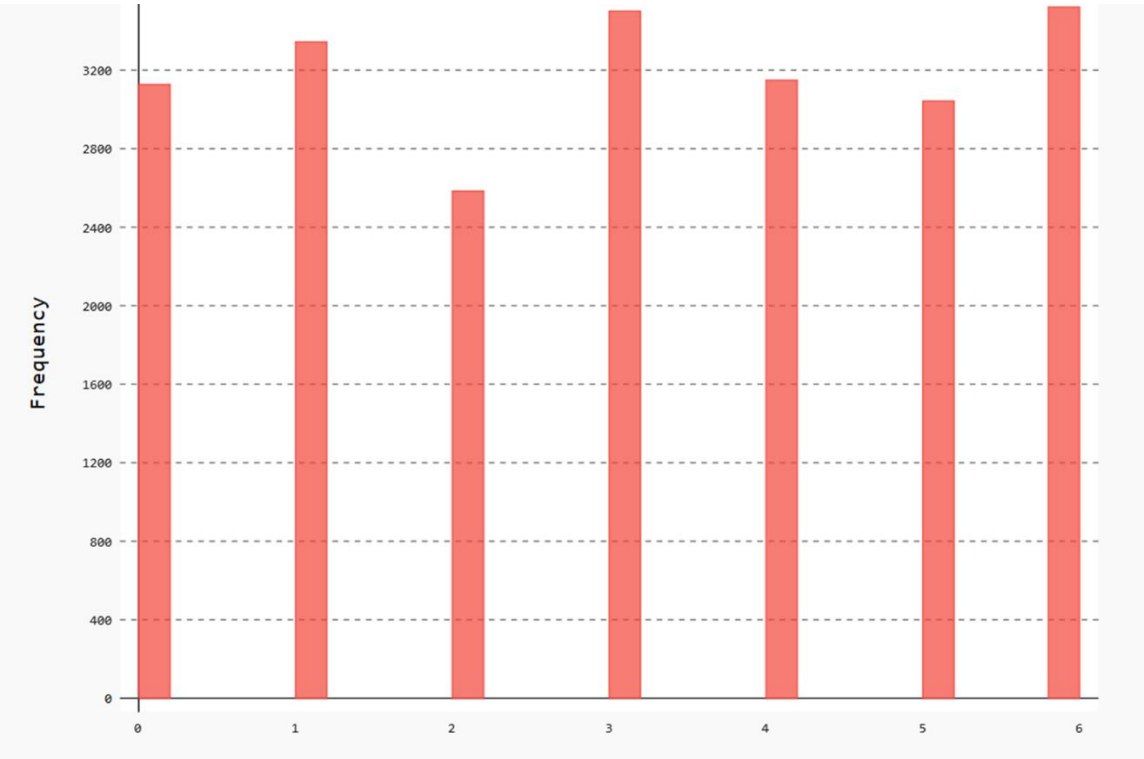

Figura 7-12 Histograma *Clusters*. Caso de Estudio 2.

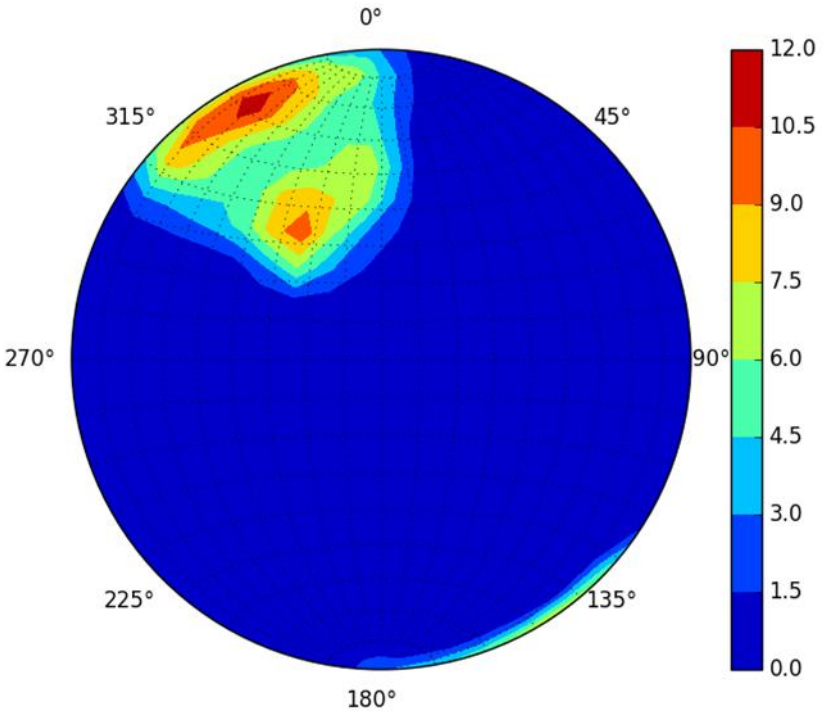

Figura 7-13 Clúster Número 1. Caso de Estudio 2.

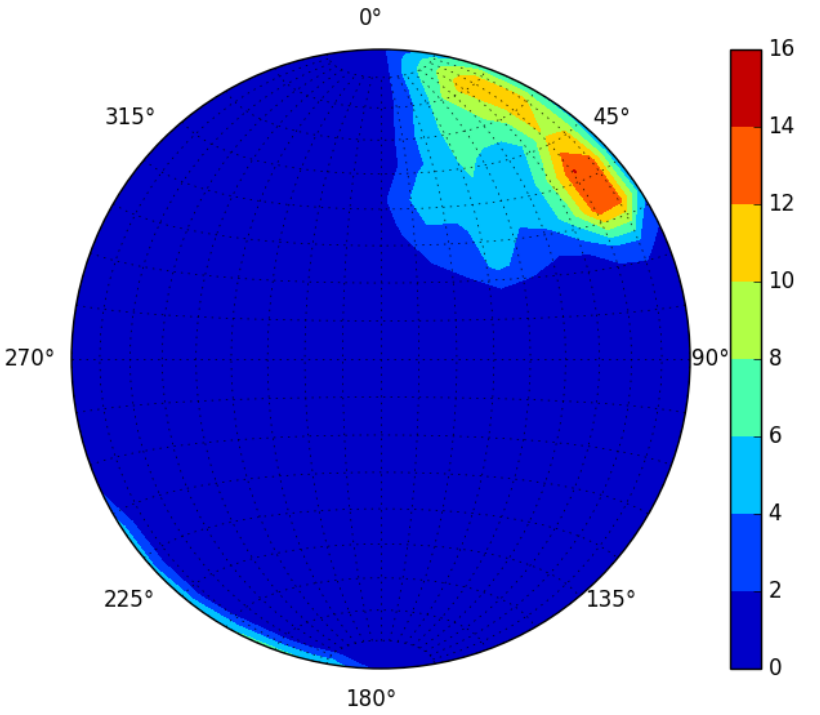

Figura 7-14 Clúster Número 2. Caso de Estudio 2.

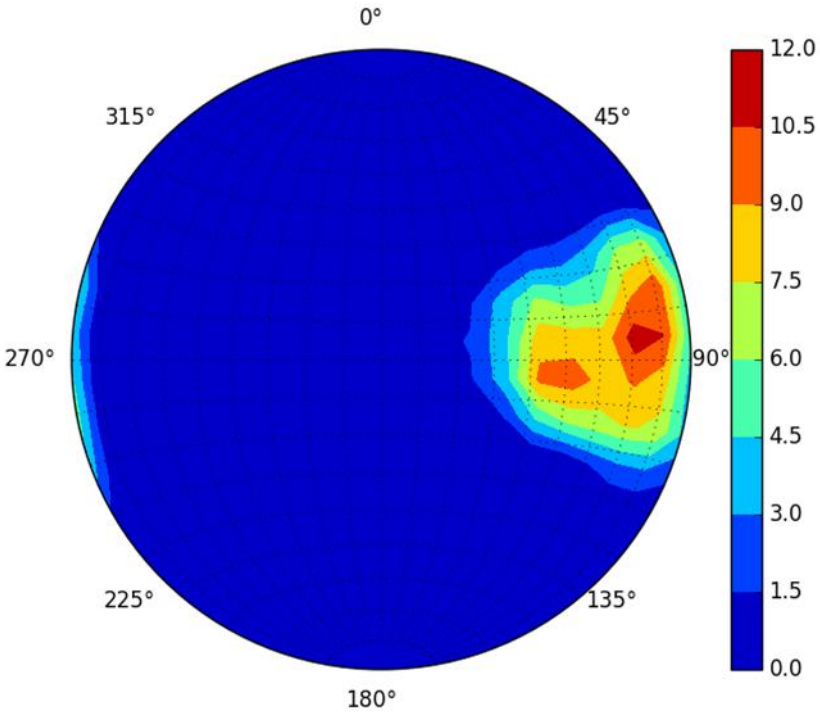

Figura 7-15 Clúster Número 3. Caso de Estudio 2.

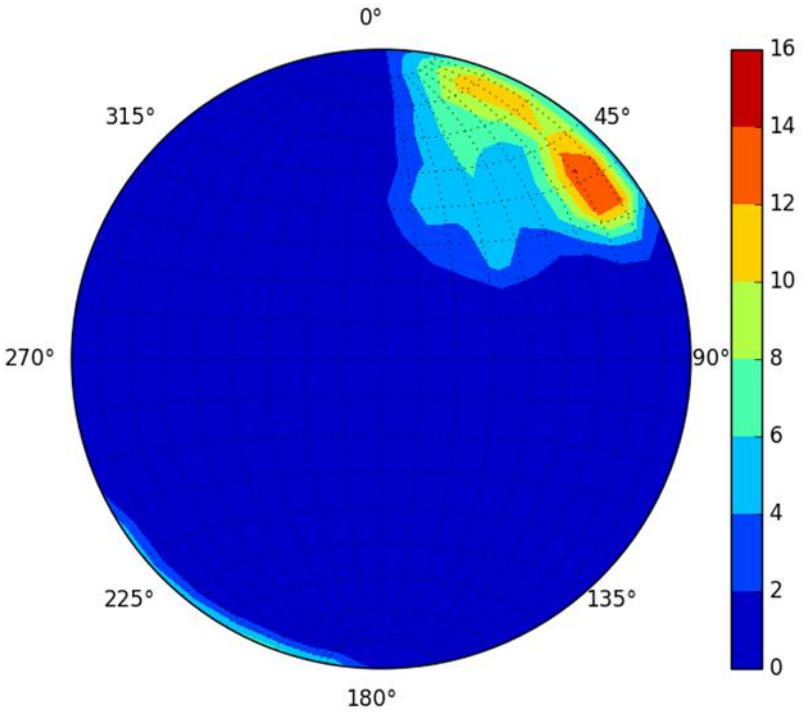

Figura 7-16 Clúster Número 4. Caso de Estudio 2.

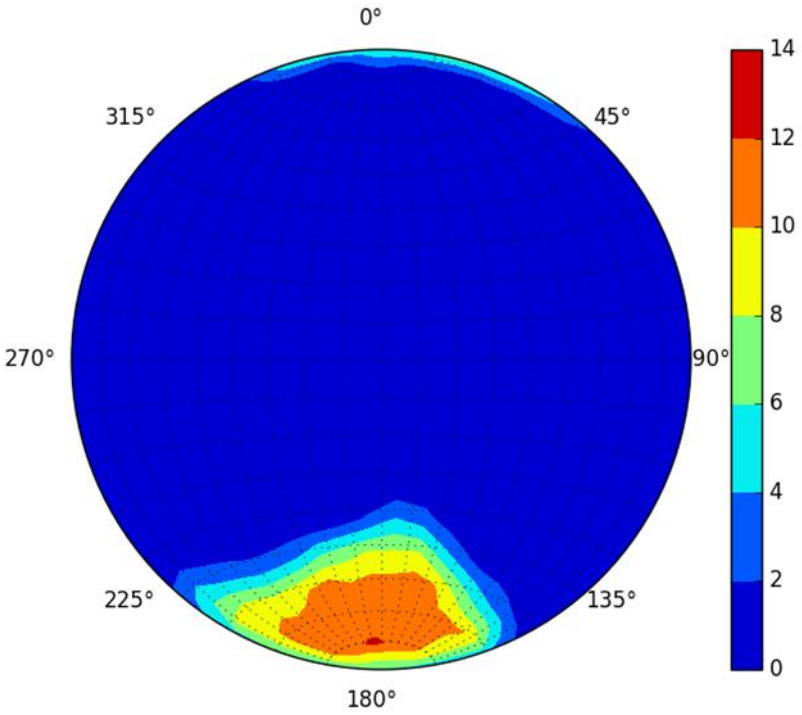

Figura 7-17 Clúster Número 5. Caso de Estudio 2.

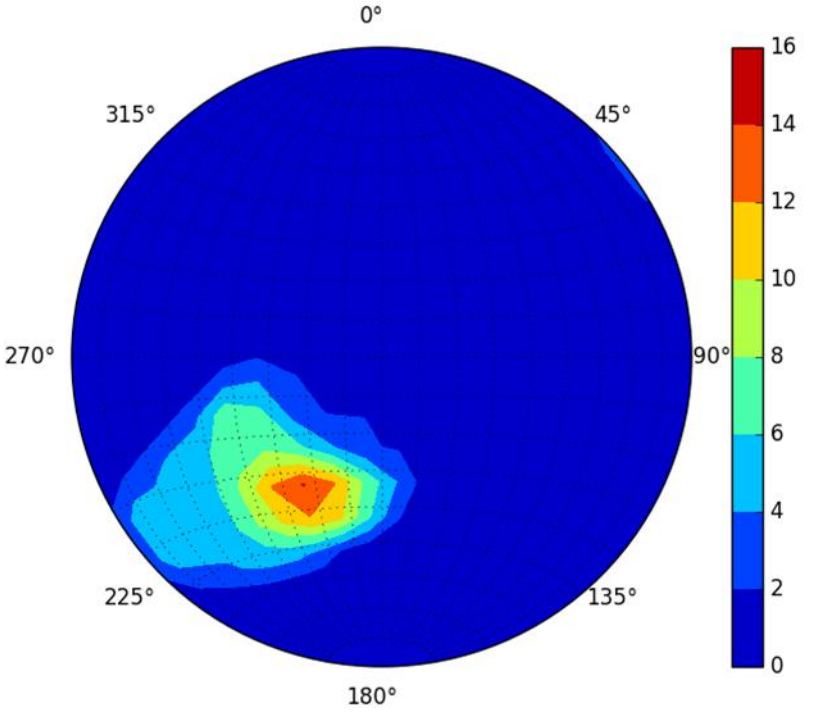

Figura 7-18 Clúster Número 6. Caso de Estudio 2.

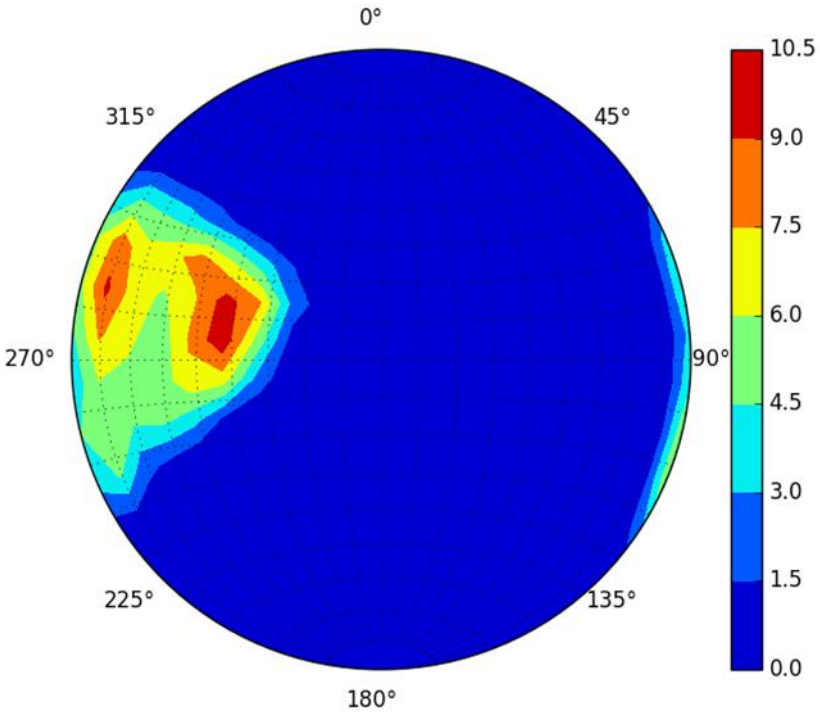

Figura 7-19 Clúster Número 7. Caso de Estudio 2.

Finalmente, se visualiza la distribución espacial de los sets. En la Figura 7-20 se observa una vista en planta de los datos. No se observa una correlación entre clúster y ubicación espacial de los datos.

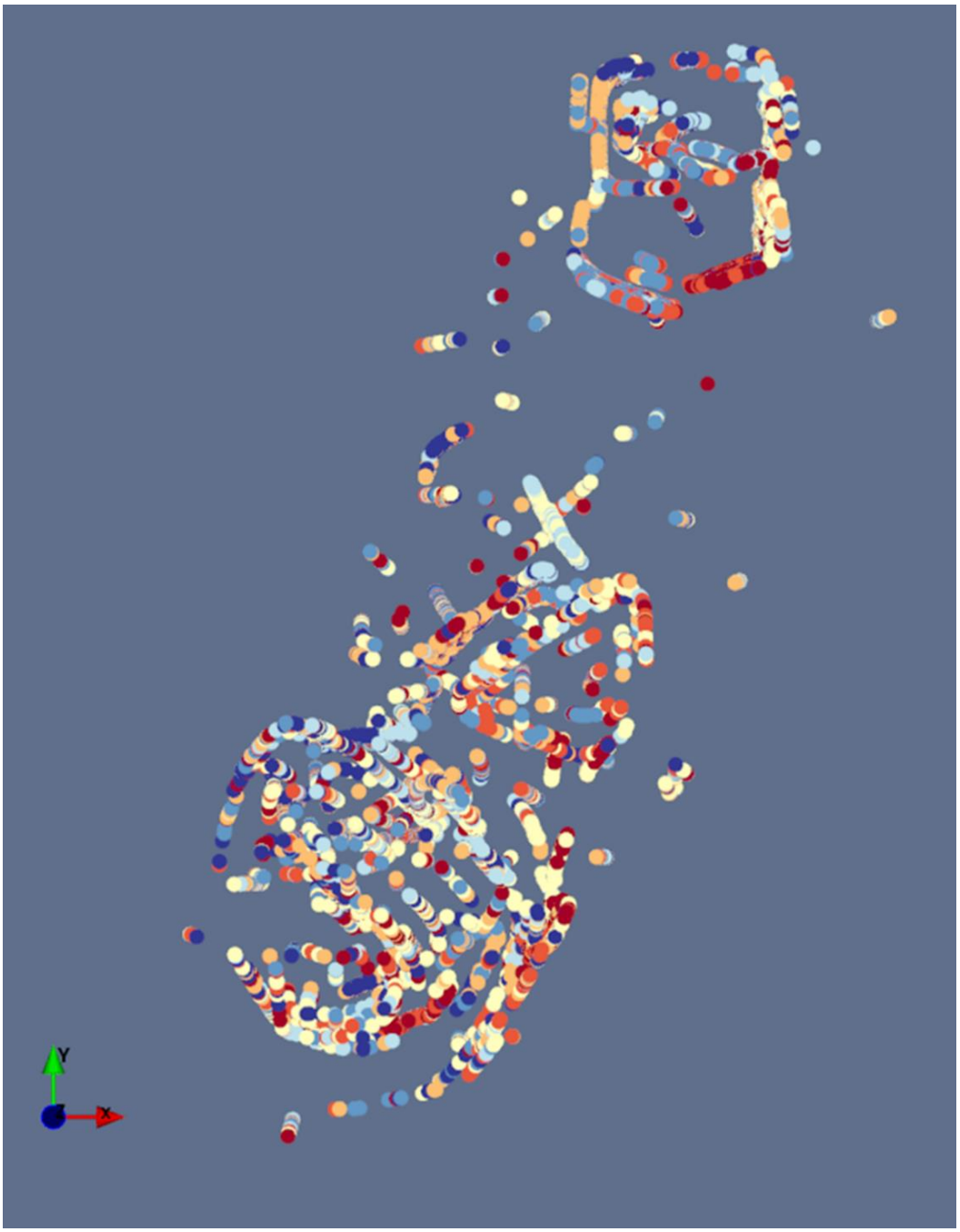

Figura 7-20 Distribución Espacial Clúster[3]. Vista en Planta. Caso de Estudio 2.

[3] Cada color de punto representa uno de los 7 clústeres identificados.

8 DISCUSIÓN

Los resultados de los algoritmos de clustering sin aplicar transformaciones muestran que ninguno de los algoritmos ha identificado los 6 conjuntos de discontinuidades esperados. La Figura 6-6, Figura 6-8, y Figura 6-10 ilustran algunos de los resultados para *K-Means, DBSCAN* y *Agglomerative Clustering*, respectivamente. Estos resultados no se ajustan a los resultados esperados, ya que los algoritmos no lograron identificar los 6 sets de discontinuidades, como se muestra en el diagrama de contorno (Figura 6-4). Estos resultados indican la necesidad de considerar transformaciones de datos para mejorar el rendimiento de los algoritmos y la precisión al momento de identificar los sets estructurales.

Los resultados con las transformaciones trigonométricas a R^2 y R^4 no logran identificar los 6 sets de discontinuidades esperados. A pesar de aplicar transformaciones a los datos, los algoritmos de clustering aún muestran un rendimiento deficiente en la identificación de los sets estructurales. Estos resultados indican que las transformaciones trigonométricas a R^2 y R^4 no son efectivas para mejorar la precisión del clustering (en este caso).

Los resultados con la transformación esférica a R^3 muestran una mejora significativa en la capacidad de los algoritmos para identificar los sets estructurales en el caso sintético. Tanto el algoritmo *K-Means* como el *Spectral Clustering*, al utilizar solo el número de clústeres como parámetro de entrada, logran reconocer los 5 clústeres en el borde junto al clúster central, lo cual se coincide con los resultados esperados.

Por otro lado, el algoritmo *Affinity Propagation* (que utiliza el parámetro de *damping* como un radio de búsqueda) obtiene una sobre-partición de los datos, identificando 7 clústeres. A pesar de esta sobre-partición, el algoritmo reconoce el clúster central, que es uno de los más difíciles de identificar.

El método del codo analiza la relación entre el número de clústers y la dispersión intraclúster. En este caso, el punto de quiebre se encuentra en 6 clústers, que coincide con lo esperado para el caso sintético (Figura 6-17).

El método de la silueta calcula la calidad de los clústers dividiendo la distancia entre los datos de un clúster y la distancia más cercana de un clúster vecino. En este caso se busca el número de clústers que maximiza este valor. En este caso el máximo se alcanza nuevamente con 6 clústers, lo que es coherente con el caso sintético (Figura 6-18).

El valor del índice de silueta es 0.48. Este índice mide cuán similares son los puntos dentro del mismo clúster y cuán diferentes son de los puntos en otros clústers. Un valor cercano a 1 indica que los clústers se encuentran bien definidos y que los puntos están más cerca de otros puntos en el mismo clúster que en otros clústers.

El valor del índice Calinski-Harabasz es 160.31. Este índice busca maximizar la relación entre la dispersión interclúster y la dispersión intraclúster. Un valor más alto indica una mejor separación entre los clústers. En este caso, el valor sugiere que la partición en 6 clústers es adecuada y que los clústers están bien diferenciados.

Por ende, las métricas indican que el algoritmo *K-Means* aplicado a los datos transformados en R^3 con 6 clústers tiene un desempeño satisfactorio. Los clústers están bien definidos y separados.

Los resultados obtenidos para el Caso de Estudio 1 indican una aplicación exitosa de la metodología propuesta. Se seleccionó el algoritmo *K-Means* con centroides inicializados en

zonas de alta densidad de polos, y divide los datos en 3 clústeres, lo cual coincide con la distribución identificada en el diagrama de contorno (Figura 7-3). Además, se observa una distribución uniforme en el histograma de los clústeres (Figura 7-4), lo que sugiere que no existe preferencia por ninguna orientación en particular en cada clúster.

Los resultados obtenidos para el Caso de Estudio 2 indican una aplicación exitosa de la metodología propuesta, al igual que en el Caso de Estudio 1. El algoritmo *K-Means*, con centroides inicializados en zonas de alta densidad de polos, divide los datos en 7 clústeres, lo que coincide con la distribución identificada en el diagrama de contorno (Figura 7-10). Similar al Caso de Estudio 1, se observa una distribución uniforme en el histograma de los clústeres (Figura 7-12), lo que sugiere que no existe preferencia por ninguna orientación en particular en cada clúster.

La metodología utilizada ha demostrado ser aplicable y efectiva en ambos casos de estudio, lo que respalda su validez y utilidad en la clasificación de datos de orientación de discontinuidades.

9 CONCLUSIONES Y RECOMENDACIONES

9.1 Conclusiones

Las principales conclusiones del trabajo realizado son las siguientes:

- ✓ Los algoritmos que solo requirieron la especificación del número de clústeres, como *K-Means* y *Spectral Clustering*, presentaron un rendimiento más efectivo y coherente en la identificación de clústeres.
- ✓ La aplicación de múltiples técnicas de selección del número de clústers, como el método del codo, la silueta y el método Gap Statistic, permitió determinar de manera efectiva el número óptimo de 6 clústers en el caso sintético.
- ✓ Los resultados reafirmaron la importancia de la selección adecuada de la transformación escogida para los datos. Esto debido a que se requiere adaptar las orientaciones de las discontinuidades a las métricas de distancia de los algoritmos. La transformación esférica a R3 se mostró efectiva, a diferencia de las otras transformaciones, como R^2 y R^4.
- ✓ Las métricas indican que los clústers obtendos con *K-Means* están bien definidos y separados, lo que es coherente con lo esperado del caso sintético.
- ✓ El algoritmo *K-Means* con inicialización de centroides basada en la densidad de datos demostró ser altamente efectivo en la clasificación de orientaciones de discontinuidades en los casos de estudio.
- ✓ La observación de distribuciones uniformes en los histogramas de clústers (Figura 7-4 y Figura 7-12) en los casos de estudio indican que no hay preferencia por orientaciones específicas en los datos geotécnicos.
- ✓ La falta de correlación entre los clústers y la ubicación espacial de los datos en los casos de estudio sugiere que las orientaciones de discontinuidades no están fuertemente influenciadas por la ubicación geográfica.
- ✓ La consistencia entre los resultados del caso sintético y los casos de estudio valida la metodología aplicada y su aplicabilidad en situaciones reales.

9.2 Recomendaciones

A continuación, se presentan algunas recomendaciones:

- ✓ Se recomienda realizar validaciones en campo de los resultados obtenidos de los casos de reales. Esto implica comparar los clústers identificados con observaciones y mediciones geotécnicas.
- ✓ A pesar de que la transformación esférica a R^3 demostró ser efectiva, se sugiere explorar otras transformaciones y métricas que puedan adaptarse aún mejor a datos de discontinuidades. Esto podría mejorar la precisión de la clasificación de clústeres.
- ✓ Se recomienda ampliar los estudios de casos reales a otras ubicaciones geográficas para a evaluar la consistencia de los resultados.
- ✓ Se recomienda comparar los resultados con los resultados que entregan herramientas de software geotécnico.

10 BIBLIOGRAFÍA

[1] A demo of K-Means clustering on the handwritten digits data. (s. f.). Recuperado 19 de julio de 2023, de https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_digits.html

[2] Andrade Fuentes, C. (2015). Levantamiento remoto de datos estructurales en rajos abiertos y su impacto en la definición de modelos estructurales: mina Los Bronces, AngloAmericano Sur.

[3] Arthur, D., & Vassilvitskii, S. (2006). k-means++: The advantages of careful seeding. Stanford.

[4] Clero, K., Ed-Diny, S., Achalhi, M., Cherkaoui, M., Benzakour, I., Soror, T., ... & Bourzeix, F. (2023). Rock mass joint set identification at Draa Sfar mine in Morocco through stereographic projection and K-means clustering. Mediterranean Geoscience Reviews, 1-8.

[5] Comaniciu, D., & Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. IEEE Transactions on pattern analysis and machine intelligence, 24(5), 603-619.

[6] Estivill-Castro, V. (2002). Why so many clustering algorithms: a position paper. ACM SIGKDD explorations newsletter, 4(1), 65-75.

[7] Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Cluster analysis 5th ed.

[8] GitHub. 2020. Joferkington/Mplstereonet. [En línea] Disponible en: https://github.com/joferkington/mplstereonet [Última consulta: 31 Julio 2023].

[9] Hammah, R. E., & Curran, J. H. (2000). Validity measures for the fuzzy cluster analysis of orientations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(12), 1467-1472.

[10] Jimenez-Rodriguez, R., & Sitar, N. (2006). A spectral method for clustering of rock discontinuity sets. International Journal of Rock Mechanics and Mining Sciences, 43(7), 1052-1061.

[11] Liu, T., Zheng, J., & Deng, J. (2021). A new iteration clustering method for rock discontinuity sets considering discontinuity trace lengths and orientations. Bulletin of Engineering Geology and the Environment, 80, 413-428.

[12] Liu, F., & Deng, Y. (2020). Determine the number of unknown targets in open world based on elbow method. IEEE Transactions on Fuzzy Systems, 29(5), 986-995.

[13] MacQueen, J. (1967, June). Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability (Vol. 1, No. 14, pp. 281-297).

[14] McClay, K. R. (2013). The mapping of geological structures. John Wiley & Sons.

[15] Mery, S. (2023). MI404-1: Análisis Estadístico y Geoestadístico de Datos, apuntes de clase. Facultad de Ciencias Físicas y Matemáticas. Universidad de Chile.

[16] Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In Advances in neural information processing systems (pp. 849-856).

[17] Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336), 846-850.

[18] Rebolledo, S. (2019). GL4102-1: Fundamentos de Geología Estructural, apuntes de clase. Facultad de Ciencias Físicas y Matemáticas. Universidad de Chile.

[19] Rocscience.com. 2020. Sets From Cluster Analysis In Dips. [En línea] Disponible en: <https://www.rocscience.com/help/dips/dips/Sets_from_Cluster_Analysis.htm> [Última consulta: 30 Julio 2023].

[20] Schubert, E., Sander, J., Ester, M., Kriegel, H. P., & Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. ACM Transactions on Database Systems (TODS), 42(3), 1-21.

[21] Scikit-learn.org. 2020. 2.3. Clustering — Scikit-Learn 0.23.1 Documentation. [En línea] Disponible en: <https://scikit-learn.org/stable/modules/clustering.html> [Última consulta: 30 Julio 2023].

[22] Sepúlveda, S (2015). CI5412-1: Mecánica de Rocas en Obras Civiles, apuntes de clase. Facultad de Ciencias Físicas y Matemáticas. Universidad de Chile.

[23] Soto, F., Hekmatnejad, A., Emery, X., & Elmo, D. (2018, November). Automatic Selection of Fracture Sets Using Clustering Techniques. In 2nd International Discrete Fracture Network Engineering Conference. American Rock Mechanics Association.

[24] Vallejos, J (2015). MI4060-1: Mecánica de Rocas, apuntes de clase. Facultad de Ciencias Físicas y Matemáticas. Universidad de Chile.

[25] Vollmer, F. W. (1995). C program for automatic contouring of spherical orientation data using a modified Kamb method. Computers & Geosciences, 21(1), 31-49.