UNIVERSIDAD EAFIT

Vigilada Mineducación

APRENDIZAJE AUTOMÁTICO PARA LA IDENTIFICACIÓN MINERALÓGICA DE MATERIAL PARTICULADO - BOGOTÁ, CALI Y VALLE DE ABURRÁ (COLOMBIA)

Machine learning for mineralogic identification of particulate matter - Bogotá, Cali and Aburrá Valley (Colombia)

Juan Alberto Gutiérrez Silva

Proyecto de grado

Asesor

Ph.D. José Fernando Duque Trujillo

UNIVERSIDAD EAFIT

ESCUELA DE CIENCIAS

GEOLOGÍA

MEDELLÍN

2023

"El cantautor y su computadora
El pastor y su afeitadora
El despertador que ya está anunciando la aurora
Y en el telescopio se demora la última estrella
La máquina la hace el hombre
Y es lo que el hombre hace con ella…"
(Jorge Drexler, 2004)

A mi madre, por apoyarme y entenderme en cuánta cosa se me ocurre desde que tengo memoria
A mi padre, porque sus enseñanzas me acompañan a donde voy
Y a mi hermano por ser mi compañero y amigo en este viaje loco de la vida

Agradecimientos

Este trabajo no sería posible sin los aportes de las siguientes personas y entidades:

Mi familia. Por brindarme lo más valioso que nadie me puede quitar: los valores y principios que me hacen ser quien soy.

Ministerio de ciencias, tecnología e innovación de Colombia y Universidad EAFIT. Por la financiación y apoyo a los programas de investigación en estimación de la polución urbana 4DAir-MOLIS.

Profesor José Fernando Duque Trujillo, jefe del Área de Territorios y Ciudades de la Universidad EAFIT. Quien no sólo me incluyó como parte del proyecto investigativo y asesoró mi trabajo. sino que también me ha brindado su apoyo, conocimiento y confianza para moverme en esto de la ciencia.

Msc. José David Avellaneda Franco e Ing. David Restrepo Rivera. Quienes ayudaron con mis dudas técnicas en el entendimiento de los datos y la programación.

Profesores Juan Darío Restrepo y Camilo Bustamante de la universidad EAFIT. que guían e instruyen con pasión el proceso de escritura académica para mí y muchos otros estudiantes.

Demás profesores de la universidad EAFIT. por sus enseñanzas y experiencias.

Mi roomie y amiga Mariana Baquero. Por facilitarme un computador cuando el mío falló en el momento más inoportuno.

Mis amigos los “lame rocas” Daniel Rodríguez, Julián Calero y Esteban de Vargas. Que siempre me ayudaron con opiniones, correcciones, respuestas a mis preguntas y sobre todo apoyo moral.

Mis amigos Santiago Díaz, Elisa Gallego, Nicolás Rodríguez, Brayan Hortúa y mención especial a Samuel Valencia por sus valiosos aportes a la redacción. El apoyo, paciencia y amistad incondicional de todos fue uno de los pilares que me mantuvieron motivado en este proceso.

Infinitas gracias siempre.

Contenido

Lista de figuras

Lista de tablas

Resumen: Identificar los componentes minerales presentes en el material particulado puede ser de gran ayuda para comprender la dinámica de la polución aérea; sobre todo para detectar la presencia de minerales peligrosos por inhalación (como los asbestos). En este trabajo se desarrolla una metodología para el agrupamiento de datos químicos obtenidos a través de microscopía electrónica de barrido con espectroscopía de energía dispersiva (SEM-EDX) en muestreos ubicados en Bogotá, Cali y Valle de Aburrá (Colombia).

Rausch et al. (2022) y Avellaneda et al. (2020) desarrollan y aplican una metodología basada en algoritmos de bosques aleatorios que permite separar categorías de partículas, entre ellas minerales. En este trabajo se desarrolla un algoritmo generalizado basado en DBSCAN como complemento. Que permitió analizar un conjunto de 3716 muestras previamente clasificadas como "mineral". Los resultados revelan la presencia de al menos 15 minerales distintos. A pesar de una efectividad de clasificación relativamente baja (~20%), este trabajo representa un avance significativo en esta área, pues los precedentes son pocos o inexistentes para este tipo de aplicaciones. Es notable, también, que se detectó la presencia de Serpentina (variedad Antigorita) en Medellín.

Las conclusiones de este estudio revelan que la mayoría de las partículas corresponden a cuarzo, calcita, caolinita y plagioclasas. A pesar de las limitaciones, el algoritmo demuestra su eficacia para la identificación de minerales. No obstante, se reconocen posibles mejoras que podrían aumentar su precisión. En conjunto, este estudio establece un punto de partida para futuros análisis de caracterización química del material particulado.

Palabras clave: Polución atmosférica, SEM-EDX, DBSCAN, caracterización química, algoritmos de agrupamiento.

Abstract: Identifying the mineral components present in particulate matter can be of great help to understand the dynamics of air pollution, especially to detect the presence of minerals that are dangerous for inhalation (such as asbestos). In this work is developed a methodology for the clustering of chemical data obtained through scanning electron microscopy with energy dispersive spectroscopy (SEM-EDX) in samples located in Bogota, Cali and Valle de Aburrá (Colombia).

Rausch et al. (2022) and Avellaneda et al. (2020) develop and apply a methodology based on random forest algorithms to separate categories of particles, including minerals. In this work, a

generalized algorithm based on DBSCAN is proposed as a complement. It allowed to analyze a set of 3716 samples previously classified as "mineral". The results reveal the presence of at least 15 different minerals. Despite a relatively low classification effectiveness (~20%), this work represents a significant advance in this area, as precedents are few or non-existent for this type of application. It is notable, also, that the presence of Serpentine (Antigorite variety) was detected in Medellín.

The findings of this study reveal that most of the particles correspond to quartz, calcite, kaolinite and plagioclase. Despite the limitations, the algorithm demonstrates its effectiveness in mineral identification. However, improvements that could increase its accuracy are recognized. Overall, this study establishes a starting point for future chemical characterization analyses of particulate material.

Key words: atmospheric pollution, SEM-EDX, DBSCAN, chemical characterization, clustering algorithms

1. Introducción:

Uno de los principales contaminantes de la atmósfera es el material particulado (PM por sus siglas en inglés), que se define como las partículas sólidas o líquidas diferentes al agua pura, que están temporalmente suspendidas en el aire (Baird y Cann, 2012). El PM puede tener un origen natural o antropogénico (Billet et al., 2018) y tener efectos nocivos sobre la salud humana, ya que es clasificado como carcinogénico (Amato, 2018). Por otra parte, la Organización Mundial de la Salud (OMS) señala a la polución atmosférica como el principal riesgo ambiental para la salud, siendo la causante de alrededor de 3 millones de muertes al año (WHO, 2016).

Respecto a la medición de estas partículas, la clasificación más usada es el índice PM, que se refiere a la concentración en µg/m³ de partículas de un tamaño dado en micrómetros. Por ejemplo, PM_{10} representa partículas inferiores a 10µm y se conoce como fracción inhalable por su capacidad de ingresar a los pulmones, mientras que $PM_{2.5}$ representa las inferiores a 2.5µm y es llamada fracción respirable, ya que por su reducido tamaño puede ingresar al torrente sanguíneo y a otros tejidos (Baird y Cann, 2012). Este trabajo se centrará en la fracción gruesa $PM_{10\text{-}2.5}$, es decir las partículas comprendidas entre 10 y 2.5µm.

En el mundo, la creciente industrialización ha incrementado la concentración y presencia del material particulado como contaminante, por lo que se han realizado diversos estudios como los de Jiang et al. (2018) en China y Pachauri et al. (2013) en India, que analizan la variabilidad temporal de la concentración y composición del PM, los riesgos para salud y también las fuentes del material. Estos trabajos encuentran correlaciones entre los cambios estacionales y las emisiones de PM por fuentes tanto naturales como antropogénicas.

En Colombia, la necesidad de monitorear la contaminación en las principales ciudades ha motivado la realización de informes de calidad del aire como los de RMCAB (2022) para Bogotá y SIATA (2021) para Medellín, centrados principalmente en la concentración y variación temporal del PM. Otros trabajos como Vargas et al. (2012) y Ramírez et al. (2018) para Bogotá; Gómez et al. (2011) y Avellaneda et al. (2020) para Medellín y Silva et al. (2020) para Barranquilla, han avanzado en caracterizar las fuentes y la química del PM, usando técnicas como la microscopía electrónica de barrido (SEM) y la espectroscopía de rayos X de energía dispersiva (EDX).

Concretamente Avellaneda et al. (2020) por medio de análisis de SEM-EDX e Inteligencia artificial de clasificación a través del algoritmo de bosques aleatorios, caracterizaron la morfo-química de

la fracción gruesa ($PM_{10-2.5}$) para el área metropolitana del Valle de Aburrá (AMVA) y diferenciaron los siguientes grupos: partículas biogénicas, partículas minerales, partículas de desgaste de llanta y partículas de abrasión metálica. Este trabajo evidencia la importancia de las fuentes antropogénicas como los vehículos, fenómenos como la re-suspensión del material particulado y el impacto de las medidas de mitigación implementadas en la zona para combatir la polución aérea.

Sin embargo, en dicho trabajo no se realiza una caracterización o identificación en grupos minerales de las partículas clasificadas como "mineral" por lo que aún hay información de importancia susceptible de ser analizada y que puede complementar el entendimiento de las posibles zonas fuente y de los efectos en la salud del PM ya que algunos minerales pueden ser nocivos, en particular los del grupo del asbesto, causantes de afecciones como el mesotelioma pleural (Berry et al., 2022). Además, la identificación de las especies minerales presentes podrían ayudar a cerrar el cerco en el análisis e identificación de fuentes de contaminación.

Teniendo en cuenta esta necesidad, la Universidad EAFIT y el Ministerio de ciencias llevan a cabo programas de investigación como 4DAir-MOLIS, que buscan desde distintos enfoques y metodologías, llenar vacíos de conocimiento en las dinámicas de la contaminación atmosférica urbana,. Este trabajo se enmarca en dicho programa investigativo y busca desarrollar una metodología semi automatizada que permita identificar los minerales presentes en la fracción gruesa del PM utilizando los datos de SEM-EDX de nuevos muestreos realizados por el programa, como complemento a la metodología mostrada por Avellaneda et al. (2020) y Rausch et al. (2022).

2. Generalidades:

2.1 Pregunta de investigación

¿Con qué grado de éxito se pueden agrupar y asociar de forma semiautomática datos composicionales provenientes de SEM-EDX que fueron clasificados como "minerales" sin identificar, con el fin de asignarles una especie mineral?

2.2 Hipótesis

Dada la variabilidad natural de la química de los minerales y las limitaciones inherentes al método SEM-EDX y al algoritmo, se prevé que es posible agrupar y asignar a una especie mineral entre un 10% y un 20% de las muestras.

2.3 Objetivo general

Desarrollar una metodología que permita el agrupamiento de datos composicionales provenientes de SEM-EDX, y previamente clasificados como "mineral" mediante la metodología de Rausch et al. (2022). Con el fin de asociarlos a especies minerales concretas.

2.4 Objetivos específicos

- Acondicionar la base de datos "ME22_Mineral.csv" obtenida durante los programas de investigación 4DAir-MOLIS como dato de salida de la metodología de Rausch et al. (2022). Y aplicada a los muestreos actuales.
- Desarrollar un programa en lenguaje Python que permita ejecutar de forma semiautomática un algoritmo de agrupamiento con los datos acondicionados, y que tenga archivos de salida tanto tabulares como gráficos para su interpretación.
- Interpretar los datos de los archivos de salida para asignar especies minerales a las distintas muestras.

3. Zona de estudio

3.1 Muestreos

Para los muestreos del programa 4DAir-MOLIS se seleccionaron las 3 ciudades más pobladas de Colombia de acuerdo con el censo del DANE para el año (2023): Bogotá, Medellín y Cali. Se cuenta con un total de 5 estaciones distribuidas de la siguiente manera: En Bogotá, la estación Centro de alto rendimiento IDRD con un muestreador a 3m del suelo que se encuentra en la zona verde de un complejo deportivo y ubicado a 500m de vías con alto flujo vehicular.

En Medellín, las estaciones Museo de Antioquia, Tanques EPM (Girardota) y Tanques la Ye. La estación Museo de Antioquia, con un muestreador a 3m del suelo está en una zona industrial que además cuenta con alto flujo vehicular y se encuentra a 235m de la línea principal del metro elevado (Avellaneda et al., 2020). La estación Girardota, con un muestreador a 2.5m del suelo. Se encuentra en una zona principalmente rural y a 10m de un depósito de escombros. Y la estación Tanques la Ye, que se encuentra en las afueras de la ciudad en un área de mucha vegetación y poco tráfico, a 500m de la vía que conduce al aeropuerto internacional José María Córdova (Avellaneda et al., 2020). Esta vía cuenta con una restricción del tráfico vehicular de carga (camiones).

En Cali, se utilizó la estación Universidad del Valle con un muestreador a 2m del suelo ubicado en una zona verde dentro de la universidad, a 80m se ubica una calle con poco tránsito y cuenta con pocos edificios cercanos.

Los puntos de muestreo fueron elegidos con miras a caracterizar zonas con distintos usos de suelo, que garantizaran la seguridad de los muestreadores, y que además contaran con estaciones meteorológicas o de medición de calidad de aire cercanas (como estaciones del SIATA para el Valle de Aburrá). El mapa de ubicación de las estaciones de muestreo se presenta en la Figura 1.

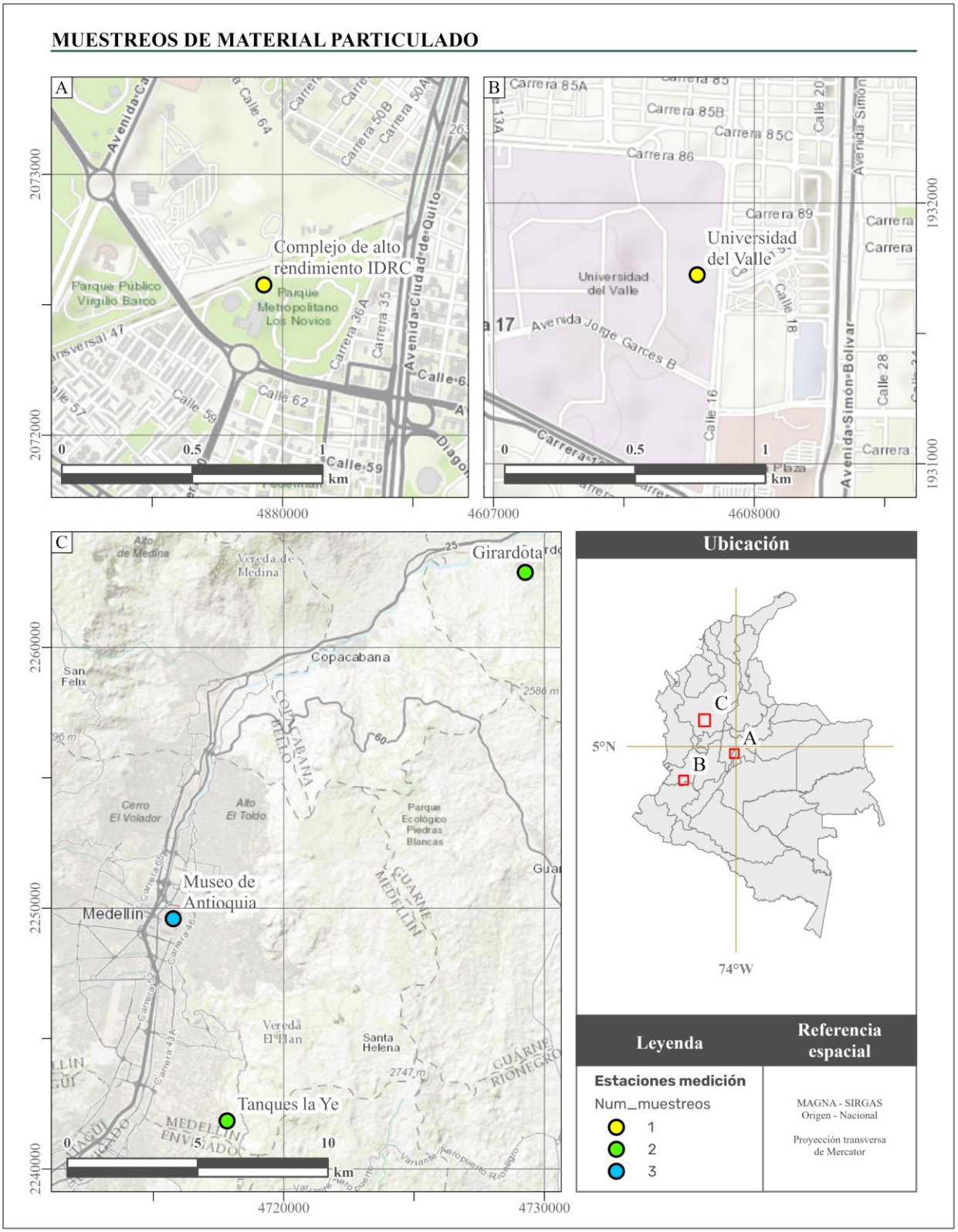


*Figura 1. Ubicación de los muestreos de $PM_{10-2.5}$. (**A**) Bogotá, (**B**) Cali, (**C**) Valle de Aburrá. Elaboración propia*

3.2 Contexto Geológico

Teniendo en cuenta que las partículas minerales presentes en el PM pueden tener origen geológico, es importante conocer la geología de las zonas muestreadas para poder generar hipótesis de los minerales que se van a encontrar. Además que en muchos casos, las rocas y minerales industriales utilizados en las ciudades son tomados de yacimientos de roca cercanos.

Bogotá: La sabana de Bogotá se enmarca en un ambiente de plegamientos y fallas en dirección NE-SW. y rocas Cretácicas y Cenozoicas levantadas durante el Mioceno-Plioceno. Las rocas más antiguas se encuentran en el núcleo de anticlinales y homoclinales, y consisten en lutitas con intercalaciones de areniscas y limolitas. Localmente se encuentran calizas marinas pertenecientes a las formaciones Chipaque-Conejo. También se encuentran areniscas, calizas y arcillolitas del Grupo Guadalupe (Cretácico marino costero). Hacia los flancos estructurales principales generados en las secuencias mencionadas anteriormente, se presentan estribaciones compuestas por rocas blandas de las Formaciones Labor y Tierna, pertenecientes al Grupo Guadalupe. (Carvajal y Navas, 2016).

En los piedemontes y valles, en contacto con fallas, se encuentran rocas sedimentarias neógenas y cenozoicas de origen marino-continental y continental, como la Formación Guaduas, con intercalaciones de arcillas y areniscas, con predominio de las areniscas en la parte media. En la parte superior de la secuencia se encuentran areniscas de grano medio a conglomerados de la Formación Cacho, que forman lomas estructurales expuestas; arcillolitas verdosas de la Formación Bogotá, con intercalaciones arenosas hacia la base; areniscas cuarcíticas de grano fino a conglomeráticas de la Formación Regadera, con intercalaciones arcillosas hacia la parte superior y Arcillolitas de color gris claro de la Formación Usme, intercaladas con areniscas de grano grueso a muy grueso en la parte superior. Cuando la Formación Usme se encuentra en núcleos de sinclinales, la erosión diferencial resulta en la inversión del relieve y se presentan crestas sinclinales dentadas y mesetas estructurales (Carvajal y Navas, 2016).

Santiago de Cali: se encuentra sobre el margen izquierdo del río Cauca, en el piedemonte oriental de la cordillera occidental, en un valle de 35km de ancho en la latitud de la ciudad. (Velásquez y Prieto, 2007).

En los cerros occidentales de Cali aflora la unidad formación Volcánica (Kv) conformada por lavas basálticas almohadilladas y columnares, diabasas y gabros. Estas rocas presentan fracturas

comúnmente rellenas con cuarzo y epidota. También pueden tener intercalaciones con rocas sedimentarias como limolitas, lodolitas y cherts. Respecto a rocas sedimentarias, se presentan las formaciones Guachinte, Jamundí y flujos de terrón colorado. La Fm. Guachinte se divide en 3 miembros: el miembro basal La Cima, que consiste en estratos gruesos de areniscas cuarzosas que van de finas a conglomeráticas intercaladas con limolitas y conglomerados. El miembro Los Chorros, con intercalaciones de areniscas, limolitas, lodolitas y shales de espesor inferior a 4 metros, este miembro es grano decreciente y presenta capas de carbón hacia el tope de la secuencia. Por último, el miembro superior La Rampla, que consiste en areniscas cuarzo feldespáticas con tamaño de grano variado y que suprayace directamente a la formación Volcánica. (INGEOMINAS Y DAGMA, 2005).

La formación Jamundí corresponde a abanicos aluviales no consolidados que afloran al suroccidente de Cali. Consiste en gravas y cantos no consolidados, pobremente seleccionados, cuyos clastos se componen de basalto, chert, gabros, limolitas, conglomerados y areniscas que van de pocos centímetros a varios metros. La matriz que envuelve a los clastos es arcillosa y rojiza. En la parte superior se encuentran niveles arcillosos y arenosos compactos y frágiles. Flujos de terrón colorado no es una formación formalmente definida, pero se incluye por su semejanza con la Fm. Jamundí. Consiste en intercalaciones de material volcánico y sedimentario que incluye fragmentos de rocas ígneas básicas (como basaltos y gabros) en una matriz limo-arenosa, y posee niveles de tobas que llegan a 1.5 metros de espesor. Se encuentra discordantemente sobre la Fm. Volcánica. Por último se encuentran los depósitos cuaternarios, que incluyen flujos de escombros, abanicos aluviales, depósito aluvial del río Cauca, depósitos aluviales activos y depósitos antrópicos. (INGEOMINAS Y DAGMA, 2005).

Valle de Aburrá: se localiza al noroccidente de la cordillera central, el Valle de Aburrá es una depresión alargada que se compone de 2 tramos. El tramo sur se encuentra entre Caldas y Bello, y el norte entre Bello y Barbosa. La zona se constituye por 3 grandes unidades litodémicas o complejos, además de cuerpos plutónicos, depósitos de vertiente y aluviales. El primero, denominado “complejo polimetamórfico de la cordillera Central” se limita por la falla Otú-Pericos al este y por el sistema de fallas de Romeral al oeste. A él pertenecen cuerpos como el ortogneis al sur del valle, cerca de la población de Caldas. Cerca a Envigado afloran gneises que varían hacia

el oriente llegando hasta granulitas y migmatitas. Sus análisis indican la presencia de varios eventos metamórficos que van del paleozoico al cretácico. (Hermelín y Rendón, 2007).

El segundo complejo, llamado Quebradagrande es una secuencia volcanosedimentaria delimitada al este por la falla de San Jerónimo y al oeste por la falla Silvia-Pijao. La parte sedimentaria se compone de cherts, grauvacas y limolitas negras, datadas para el cretácico temprano. La parte volcánica comprende tobas, aglomerados, andesitas y basaltos con edades también del cretácico temprano. En tercer lugar, los complejos ofiolíticos caracterizados por su alta deformación y límites fallados norte sur. En ellos afloran dunitas, peridotitas, gabros y metagabros. Existen varios modelos de emplazamiento para estas unidades, y sus edades radiométricas van del Jurásico tardío al cretácico tardío. (Hermelín y Rendón, 2007).

Los cuerpos plutónicos presentes en la zona se caracterizan por su variación composicional, edad cretácica, relaciones intrusivas y afectaciones por fallamientos, que indican una actividad magmática posterior a la acreción de los cuerpos metamórficos, al igual que una gran actividad tectónica. Los cuerpos destacados son el Batolito Antioqueño y los stocks de Ovejas, de las Estancias, de Altavista, y de San Diego. Con composiciones que van de gabros a granodioritas. Por último, los depósitos no consolidados que representan más del 50% del área del valle, entre los que se encuentran depósitos aluviales, aluviotorrenciales, coluviales y flujos de escombros/lodos. Los aluviones corresponden a los depósitos del río Medellín y sus afluentes y poseen espesores variables que llegan a superar los 200m. (Hermelín y Rendón, 2007).

4. Marco conceptual

De acuerdo con la definición presentada en Dana (1864). Una de las características clave que determinan a los minerales es su composición química. en muchos casos, conocer la composición, fórmula química o porcentaje de elementos de alguno puede directamente derivar en su identificación con excepciones como los minerales polimórficos.

Una forma de obtener datos químicos es a través de la espectroscopía de rayos X de energía dispersiva (EDX por sus siglas en inglés). Este método analiza los picos de espectro de rayos X generados por las interacciones electrón-átomo durante un análisis de microscopía electrónica de barrido, y proporciona una cuantificación de los elementos químicos presentes en la muestra (Goldstein et al., 2017).

Dado el gran volumen de datos (más 3700 muestras clasificadas como “mineral”). Es pertinente un enfoque que permita analizarlos de forma conjunta. De manera viable en términos de tiempo.

Para esta tarea existen herramientas como los algoritmos de agrupamiento, procedimientos estadístico-matemáticos que permiten separar datos en grupos de acuerdo con algún criterio, que generalmente se relaciona con el concepto de similitud o proximidad entre datos (Xu y Wunsch, 2005). El algoritmo necesario para este proyecto debe cumplir con los siguientes criterios:

- Poder agrupar los datos sin necesidad de indicar manualmente el número de “clústeres” o grupos. Ya que la cantidad de especies minerales presentes en los muestreos es desconocida.
- Poder controlar de alguna manera la exclusión de datos ruidosos o atípicos. Debido a la variación de la química mineral y las posibles fallas de precisión del método SEM-EDX.
- Ser de relativamente baja complejidad con el fin de integrarse en los requerimientos de tiempo y extensión de un proyecto de pregrado**.**

El algoritmo seleccionado fue el agrupamiento espacial basado en densidad de aplicaciones con ruido “DBSCAN”, propuesto por Ester et al. (1996). Este se encuentra integrado a la librería de código abierto *Scikit-learn* para el lenguaje de programación *Python*.

DBSCAN es catalogado como aprendizaje no supervisado, y funciona ubicando los datos como puntos en un espacio de tantas dimensiones como variables se tengan. Posteriormente se establece una distancia euclidiana o radio máximo denominado Épsilon (ε), hasta el que se consideran los puntos como parte de un mismo grupo o clúster. Además se establece un número mínimo de puntos que deben estar dentro de ese radio para considerarse como un grupo. El tamaño, forma y número de los grupos dependerán de épsilon y del mínimo de puntos. (Ester et al., 1996). En la Figura 2 se ilustran los parámetros mencionados.

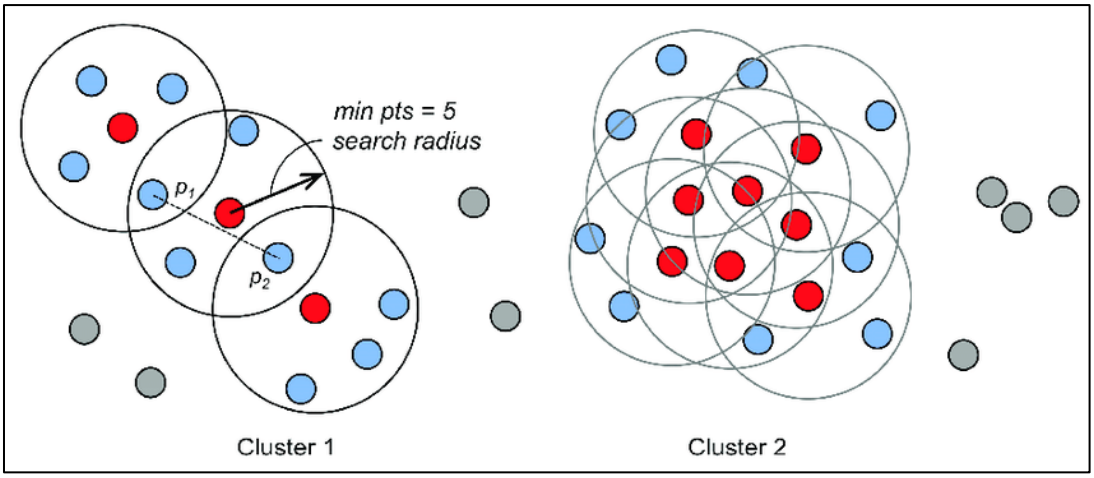


Figura 2. Épsilon (radio de búsqueda) y mínimo de puntos para DBSCAN. Los puntos grises son identificados como ruido. Tomado de DiFrancesco Et al. (2020)

El valor épsilon y el número de puntos pueden elegirse arbitrariamente o con ayuda de métodos que permitan entender la naturaleza de los datos, como diagramas de puntos cercanos vs. Distancia. (Ester et al., 1996). A estos diagramas se les ubica un punto óptimo de curvatura que corresponde en una de sus coordenadas a la distancia o épsilon óptimos. A esto se le conoce como método del codo.

5. Metodología:

Este trabajo de grado se enmarca en el programa de investigación MinCiencias (2021-2023): Estimación de la polución urbana mediante el uso de mediciones y asimilación de datos en superficie, in situ y de detección remota (4DAir-MOLIS). Que busca contribuir al conocimiento para la modelación y entendimiento de la contaminación atmosférica.

5.1 Recolección de muestras y toma de datos

La toma de muestras del programa se realizó en los puntos previamente mencionados. Mediante los muestreadores pasivos modelo Sigma-2 (no requieren bombeo de aire) desarrollados por la empresa Particle Vision inc. (Anexo 1). Estos cuentan con un sustrato de boro y una lámina adhesiva que los hace adecuados para este tipo de análisis (Avellaneda et al., 2020). Los muestreos se encuentran resumidos en la Tabla 1.

Tabla 1. Resumen de los muestreos.

Ciudad	Estación	Muestreo	Fecha inicio	Fecha fin	Total partículas muestreadas
Bogotá	Centro de alto rendimiento IDRD	BO22_Car_p1	17/02/2022	24/02/2022	892
Cali	Universidad del Valle	CA22_UV_p1	8/03/2022	17/03/2022	871
Medellín	Museo de Antioquia	ME22_MA_P1	21/04/2022	29/04/2022	1218
Medellín	Museo de Antioquia	ME22_MA_P2	3/11/2021	15/11/2021	1023
Medellín	Museo de Antioquia	ME22_MA_P3	7/02/2022	15/02/2022	851
Medellín	Tanques Girardota	ME22_TG_P1	3/11/2021	15/11/2021	333
Medellín	Tanques Girardota	ME22_TG_P2	5/10/2021	8/10/2021	984
Medellín	Tanques la Ye	ME22_TLY_P1	3/11/2021	15/11/2021	286
Medellín	Tanques la Ye	ME22_TLY_P2	5/10/2021	8/10/2021	962
					7420

Posteriormente se realizó la toma de datos SEM-EDX utilizando un umbral de grises homogéneo para todas las partículas. El microscopio utilizado fue el ZEISS GEMINI SEM 300 acoplado con un EDX Oxford XMAX. Se utilizó una ventana de 80mm^2, un voltaje de aceleración de 12.000V y una magnificación de 500X. al ser el elemento del sustrato, la señal de boro fue retirada de todas las muestras (Avellaneda et al., 2020).

Los datos obtenidos se clasificaron con inteligencia artificial mediante el algoritmo supervisado de bosques aleatorios desarrollado por Particle Vision Inc. (2020). Ilustrado en Avellaneda et al. (2020) y Rausch et al. (2022) La definición de los grupos morfoquímicos y de los parámetros morfométricos calculados para esta clasificación se encuentran en los Anexo 60 y Anexo 61. El resumen de estos resultados se presenta en la Figura 3.

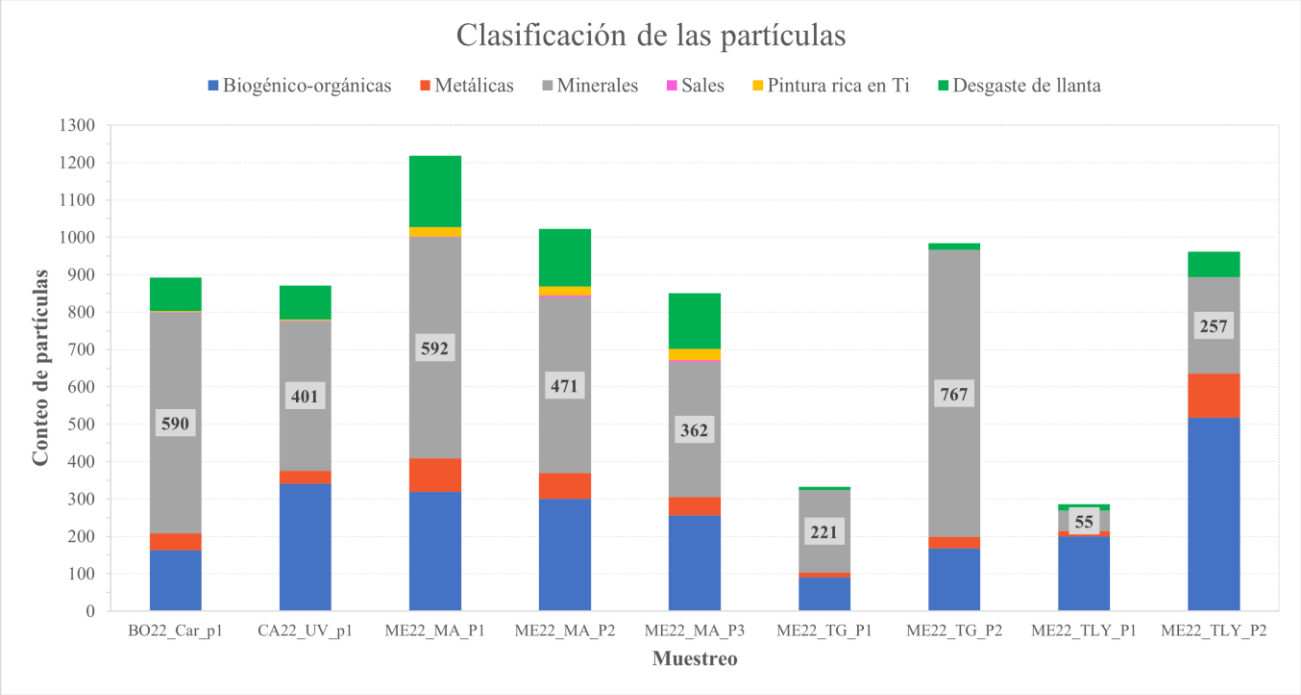


Figura 3. Resumen de la distribución de los distintos grupos morfoquímicos en los muestreos, énfasis en "minerales" con un total de 3716 partículas. Elaboración propia.

5.2 Acondicionamiento de los datos

Para el análisis correspondiente a este trabajo se utilizó la base de datos de las partículas clasificadas como "mineral" de los 9 muestreos mencionados. Un total de 3716 muestras con datos químicos de EDX, expresados como porcentaje en peso por elemento.

Como acondicionamiento inicial se eliminaron de la base de datos las variables no correspondientes a elementos químicos (parámetros morfométricos), al igual que las variables de elementos cuyo valor para todas las muestras fuese cero, quedando con un total de 36 variables (elementos) por muestra. Los elementos químicos eliminados y conservados de la base de datos se ilustran en el Anexo 2.

A la base de datos se agregó una "biblioteca" o lista de 42 minerales conocidos con su porcentaje en peso de elementos. Con el fin de que sirvan como etiquetas a la hora de realizar el agrupamiento. Dichos porcentajes fueron tomados de la base de datos de Mineralogy Database (https://www.webmineral.com). Esta lista de minerales se muestra en el Anexo 62.

5.3 Clustering mediante algoritmo de aprendizaje no supervisado DBSCAN.

El código para la ejecución del algoritmo de agrupamiento DBSCAN fue escrito en el entorno *Jupyter Notebook* para el lenguaje *Python*. Cuenta con cuadros de diálogo para ingresar datos, variables y parámetros. Ofrece un gráfico de puntos cercanos vs distancia para la estimación de Épsilon y salidas de clústeres en formato CSV con gráficos de líneas en formato PNG. Los clústeres de salida se agregan como una variable más a la base de datos de entrada, y se numeran a partir de 0, siendo el "clúster -1" correspondiente a los datos identificados como ruido.

El diseño "intuitivo" se hizo para facilitar y generalizar el uso de este código por parte de personas no familiarizadas con la programación, o que deseen darle otras aplicaciones.

El código desarrollado se encuentra en el Anexo 64. El notebook para *Jupyter* listo para su descarga se encuentra publicado a través de *GitHub* con acceso libre. Se puede acceder al repositorio en el siguiente enlace:

https://github.com/Gutierrez-Juan/DBSCAN.git

El algoritmo se ejecutó con los datos tomados para cada ciudad y se repitió de forma iterativa cuando alguno de los clústeres de salida tuviera una cantidad elevada de datos con variabilidad evidente, con el fin de separarlo en más grupos. Las muestras de dichos clústeres se toman como nuevos datos de entrada para el agrupamiento.

5.4 Identificación.

Para la identificación mineral se tomaron los resultados del algoritmo. Las muestras en cuyo clúster se incluyó alguno de los minerales etiquetados, se asignan con alta probabilidad a ese mismo mineral.

En otros casos, el algoritmo asigna clústeres sin algún mineral etiquetado, por lo que su identificación se realiza analizando manualmente los porcentajes de los elementos mostrados, con apoyo de herramientas como la búsqueda inversa por elementos de Mineralogy Database (https://www.webmineral.com) y de las microfotografías SEM.

6. Resultados

Debido a su cantidad, los gráficos de los clústeres obtenidos a partir del código se muestran en el apartado de anexos de la siguiente forma: del Anexo 3 al 9 para Bogotá. Del Anexo 10 al 21 para Cali y del Anexo 22 al 59 para el Valle de Aburrá. Los resultados generales se resumen en la Figura 4 y 5.

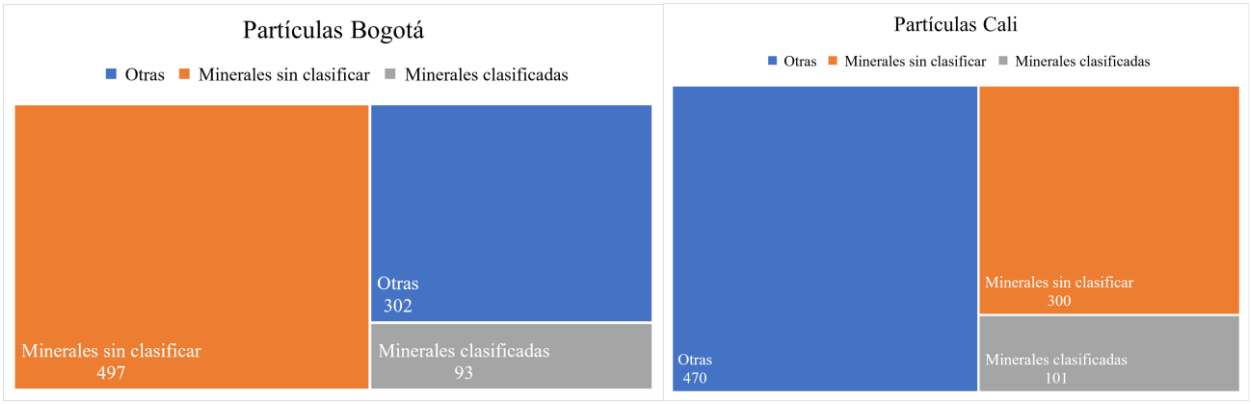


Figura 4. Resumen de la clasificación de partículas. Bogotá y Cali. Elaboración propia

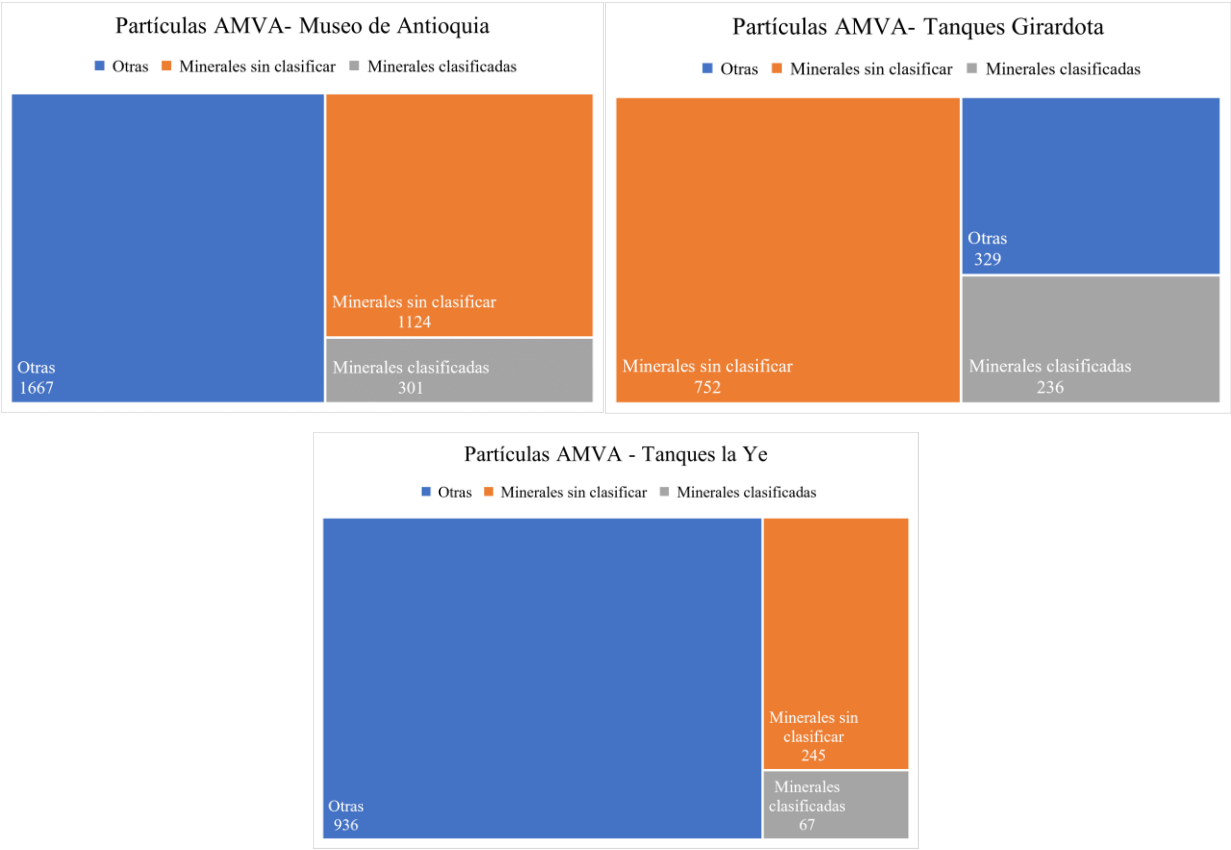


Figura 5. Resumen de la clasificación de partículas. Valle de Aburrá. Elaboración propia

Dados los parámetros épsilon y mínimo de puntos configurados en cada iteración del algoritmo (mostrados en el Anexo 63). Una gran cantidad de los datos fue identificada como ruido, aunque los grupos conservaron baja variabilidad. El porcentaje de muestras identificables con este método osciló entre un 15.7% y un 25.1% del total de muestras minerales.

7. Discusión

En la Tabla 2. se detalla la interpretación dada a cada uno de los clústeres resultantes del algoritmo para las tres ciudades. Los nombres de los clústeres se abrevian de la siguiente forma: "*Iniciales de la zona* – I(*iteración*) – Cl-(*número de clúster*)". Cabe destacar que los minerales que se pueden asociar a las muestras no mencionadas en la tabla siguen siendo desconocidos. Por lo que en total podrían ser cantidades diferentes de los mismos, o puede que aún falten minerales por identificar.

Tabla 2. Clústeres con su interpretación. Elaboración propia.

Cluster	# Figura Anexa	Mineral etiquetado	# Muestras en el cluster	Descripción/interpretación
Bog-I1-Cl-0	3	-	354	Usado para la iteración 2
Bog-I1-Cl-1	4	**Calcita**	9	Leves variaciones de Al, Si, S y Fe
Bog-I1-Cl-2	5	-	5	Plagioclasa con variación en Fe
Bog-I1-Cl-3	6	-	4	Calcita con señal mezclada de Al, Si y Fe
Bog-I2-Cl-0	7	**Cuarzo**	50	Leves variaciones de Al, K, Ca y Fe
Bog-I2-Cl-1	8	-	12	Cuarzo con variación anómala de C
Bog-I2-Cl-2	9	**Caolinita**	16	Variación de Ca
Cal-I1-Cl-0	10	-	9	Cuarzo con variación anómala de C
Cal-I1-Cl-1	11	**Albita**	6	Variación de Ca y Fe
Cal-I1-Cl-2	12	-	20	Plagioclasa con variación en C y Fe
Cal-I1-Cl-3	13	-	135	Usado para la iteración 2
Cal-I1-Cl-4	14	-	11	Moscovita con variación en Fe
Cal-I1-Cl-5	15	-	4	Posiblemente hornblenda con anomalía en C
Cal-I1-Cl-6	16	-	5	Aluminosilicato indeterminado, posiblemente arcilla. Anomalía en C
Cal-I1-Cl-7	17	-	7	Calcita con señal mezclada de Al, Si y Fe
Cal-I1-Cl-8	18	**Calcita**	8	Leves variaciones de Al, Si, S y Fe
Cal-I2-Cl-0	19	-	10	Hornblenda con anomalía de C
Cal-I2-Cl-1	20	-	14	Hornblenda, más férrica con anomalía de C
Cal-I2-Cl-2	21	-	7	Hornblenda con anomalía de C
AMVA-I1-Cl-0	22	-	1336	Usado para la iteración 2
AMVA-I1-Cl-1	23	-	21	Actinolita-tremolita con anomalía en C
AMVA-I1-Cl-2	24	-	10	Olivino con anomalía de C
AMVA-I1-Cl-3	25	-	38	Olivino con mayor anomalía de C
AMVA-I1-Cl-4	26	-	23	Yeso con anomalía de C
AMVA-I1-Cl-5	27	**Calcita**	43	Leves variaciones de Al, Si, S y Fe
AMVA-I1-Cl-6	28	-	9	Olivino con anomalías de Ca y C
AMVA-I1-Cl-7	29	-	23	Calcita con señal mezclada de Al, Si y Fe
AMVA-I1-Cl-8	30	-	5	Aluminosilicato de Fe y Na indeterminado
AMVA-I1-Cl-9	31	-	11	Error de toma de datos EDX (muestra 100% de oxígeno)
AMVA-I1-Cl-10	32	-	5	Aluminosilicato de Fe y Ca indeterminado (posiblemente epidota)
AMVA-I1-Cl-11	33	-	6	Aluminosilicato de Fe y K indeterminado
AMVA-I1-Cl-12	34	-	5	Calcita con señal mezclada de Al, Si y Fe
AMVA-I1-Cl-13	35	**Antigorita**	6	Familia de las serpentinas, se procede a verifiicar en imágenes SEM
AMVA-I1-Cl-14	36	**Ortoclasa**	5	Variación mínima en Na
AMVA-I1-Cl-15	37	-	5	Señal mezclada entre aluminosilicato de Fe y carbonato
AMVA-I1-Cl-16	38	-	5	Yeso con anomalía de C
AMVA-I2-Cl-0	39	**Cuarzo**	84	Leves variaciones de Al, K, Ca y Fe
AMVA-I2-Cl-1	40	-	6	Aluminosilicato de Ca y Fe (posiblemente epidota). fuerte anomalía de C
AMVA-I2-Cl-2	41	**Albita**	67	Variación de Ca, Fe y C
AMVA-I2-Cl-3	42	-	20	Turmalina (chorlo)
AMVA-I2-Cl-4	43	-	8	Aluminosilicato de Fe y K indeterminado
AMVA-I2-Cl-5	44	-	5	Óxido de hierro con señales mezcladas de C, Mg, Al y Si
AMVA-I2-Cl-6	45	-	52	Turmalina (chorlo) con anomalía de C
AMVA-I2-Cl-7	46	-	19	Biotita
AMVA-I2-Cl-8	47	**Caolinita**	42	Variación de Ca y K
AMVA-I2-Cl-9	48	-	5	Posiblemente biotita con anomalía en C
AMVA-I2-Cl-10	49	-	28	Señal mezclada entre aluminosilicato, Fe y C
AMVA-I2-Cl-11	50	-	10	Aluminosilicato de Fe, Ca y Mg indeterminado (posiblemente piroxeno)
AMVA-I2-Cl-12	51	-	5	Señal mezclada entre aluminosilicato de Fe y Mg (Px) y óxido de hierro
AMVA-I2-Cl-13	52	-	4	Se interpreta como osumilita (por medio de la búsqueda de webmineral)
AMVA-I2-Cl-14	53	-	8	Moscovita
AMVA-I2-Cl-15	54	-	10	Señal mezclada entre aluminosilicato, óxido de hierro y Carbono
AMVA-I2-Cl-16	55	-	6	Señal mezclada entre aluminosilicato, óxido de hierro y Carbono
AMVA-I2-Cl-17	56	-	5	Señal mezclada entre cuarzo y aluminosilicato de Ca y Fe.
AMVA-I2-Cl-18	57	-	5	Señal mezclada entre aluminosilicato, óxido de hierro y Carbono
AMVA-I2-Cl-19	58	-	6	Óxido de hierro, con señal mezclada de aluminosilicato de Mg
AMVA-I2-Cl-20	59	-	6	Óxido de hierro, con señal mezclada de aluminosilicato

Es importante señalar que no fue posible asignar un mineral con total certeza a todos los clústeres. Pero sí para la mayoría, especialmente aquellos que incluyeron alguna de las etiquetas de minerales. Se destaca la aparición repetitiva de cuarzo, calcita, albita y caolinita.

Se puede decir que todos los minerales asignados a los clústeres son coherentes con la geología de sus zonas respectivas. Un posible origen antrópico de las partículas también explicaría su detección, ya que muchas de ellas corresponden a minerales que se usan comúnmente en la industria de la construcción o que están presentes en procesos de explotación minera. También se debe tener en cuenta que la abrasión del material de pavimento podría ser una posible fuente de estas partículas. Esto se sospecha por la anomalía de carbono que se presenta de forma reiterada en las muestras y que puede ser causada por el desgaste de carreteras asfaltadas. En este caso podría esperarse que partículas de brea del pavimento y otros materiales con carbono queden adheridos a las partículas suspendidas. De ser cierta esta hipótesis, podría afirmarse que la mayor fuente de minerales en el material particulado de este análisis corresponde a la abrasión causada por los vehículos en las carreteras y no directamente de la geología, o por lo menos para las partículas identificadas

La baja cantidad de muestras identificadas puede generar un sesgo en la información y obstaculizar una medición objetiva de la abundancia relativa de cada mineral. La metodología debe avanzar y mejorar para identificar más datos y así poder realizar este tipo de cálculos. Por tanto, el alcance actual se ve limitado a confirmar la presencia de ciertos minerales en el PM.

Para el caso del Valle de Aburrá, se destaca notablemente el clúster AMVA-I1-Cl-13, que fue agrupado junto con la etiqueta para antigorita. La antigorita es un mineral perteneciente a la familia de las serpentinas, y se forma sólo por encima de los 250°C (Wenner & Taylor, 1974). La detección de este mineral es importante dado a que pertenece a una familia de minerales conocidos por sus efectos nocivos en la salud respiratoria humana (Gwenzi, 2020). Si bien la cantidad de partículas incluida en el clúster es baja, se hace pertinente una corroboración por medio de las imágenes SEM, ya que de confirmarse puede ser un punto de partida para futuros estudios. En la Figura 6 se muestran las microfotografías SEM de las muestras del clúster AMVA-I1-Cl-13.

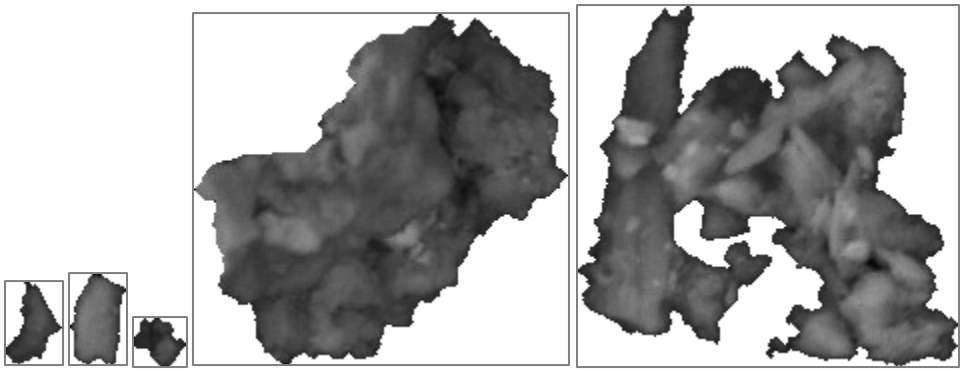


Figura 6. Microfotografías SEM de las muestras asignadas como Antigorita. de izquierda a derecha 1) ME22_MA_P1_1289; 2) ME22_TG_P1_345; 3) ME22_TG_P2_1238; 4) ME22_TG_P2_1880 y 5) ME22_TLY_P2_1340.

Debido a la resolución de las imágenes, sólo es posible hacer una distinción clara de las características de las partículas 4 y 5. Siendo la 5 la que presenta un hábito astilloso y formas alargadas más marcadas. Características diagnósticas de la antigorita. Esta partícula fue recolectada en la estación Tanques la Ye.

Una de las mayores dificultades para el método desarrollado radica en que la mayoría de los clústeres presentan variaciones de elementos que no pertenecen a la composición ideal del mineral asignado. Estas variaciones pueden corresponder a diferentes causas, entre ellas: 1- Variación natural de la química mineral. Ya que el método de clustering pretende de cierta manera discretizar características que muchas veces hacen parte de un continuo. Como la química de los minerales en soluciones sólidas. 2 – impurezas. Las impurezas pueden encontrarse naturalmente en los minerales, la anomalía de carbono mencionada anteriormente dificulta en gran medida la identificación mineral al enmascarar valores de otros elementos que podrían ser diagnósticos. 3 - errores de detección del EDX. En el que se destaca el caso del clúster AMVA-I1-Cl-9, que fue correctamente agrupado por el algoritmo, pero todos sus datos indican una composición de 100% oxígeno para algo clasificado previamente como mineral, lo que evidencia que el método EDX puede tener falencias en la toma de sus datos por falta de calibración, interferencias, etc. 4 – Señales mezcladas. Debido a la amplitud del haz de electrones del SEM es posible que los datos obtenidos por el EDX estén caracterizando a más de un cristal o mineral a la vez. También es posible que algunas partículas individuales estén compuestas por varios minerales. Por último la

meteorización/alteración de los minerales puede influir en los datos, esto se evidencia principalmente por la aparición de valores altos de hierro y oxígeno en clústeres de minerales ferromagnesianos (como los piroxenos).

Es innegable que el algoritmo aun presenta una baja capacidad de agrupamiento preciso para muchos datos, siendo capaz de separar alrededor de un 20% de las muestras ingresadas con los parámetros épsilon y min_pts suministrados. Este resultado es esperable dada la naturaleza compleja y variada de las partículas, además de la naturaleza pionera de este estudio.

En cuanto a las limitaciones inherentes al algoritmo se destaca que la elección de un valor óptimo de épsilon se hace más difícil entre más variables se tengan. ya que un valor que pueda distinguir correctamente los grupos de una variable puede ser insuficiente o exagerada para otras. En este caso se intentó resolver este problema por medio de una normalización de los datos, que uniformiza las escalas de las diferentes variables. Sin embargo, este proceso parece ser insuficiente dentro del presente trabajo. Métodos como el "método del codo", cuyo objetivo es hallar este valor óptimo no son efectivos para arreglos de datos con muchas dimensiones (o variables).

Finalmente para el presente trabajo se optó por valores de épsilon que formaran grupos de gran similitud entre datos a costa de disminuir la cantidad de estos. Con el fin de realizar una identificación confiable basada en la composición química.

La presencia de datos ruidosos presenta otra dificultad, ya que el ruido puede ser consecuencia de la propia naturaleza del material, por lo que su identificación, aun con un mejoramiento de los algoritmos depende meramente de una identificación particular (que puede ser costosa computacionalmente si se tienen muchos datos). Por tanto, el análisis puede reducirse por ahora a intentar identificar una fracción de las muestras que se considere estadísticamente significativa, hasta que se pueda llegar a una identificación refinada.

La principal limitación del método radica en que únicamente está teniendo en cuenta la composición química para intentar diferenciar unos minerales de otros, por lo que el acoplamiento de este junto con otros métodos que puedan diferenciar y asociar morfologías a partir de las imágenes SEM o incluso estructuras por medio de otros métodos (como la difracción de rayos X) podrían mejorar notablemente la capacidad y confiabilidad de este tipo de estudios a futuro.

Una forma sencilla de facilitar la identificación mineral con este algoritmo podría ser incrementando el volumen de la biblioteca de espectros. Bien sea con datos publicados, o mediante la toma experimental de nuevos datos que se adapten a las necesidades del análisis.

Este trabajo sienta un precedente en cuanto a la identificación de los minerales presentes en el PM de las ciudades de Colombia. Se espera que a futuro la realización de este y otros análisis contribuyan significativamente a la comprensión de la dinámica de las partículas contaminantes dada su importancia para la salud pública.

Por otra parte, el desarrollo del algoritmo de manera generalizada es también un punto de partida para la realización de análisis de otras índoles. El notebook desarrollado en este trabajo para DBSCAN está en capacidad de agrupar datos numéricos de cualquier tipo y de un número personalizable de variables. De una forma relativamente intuitiva y fácil de usar incluso por personas ajenas a la programación, por lo que puede convertirse en una herramienta de acceso libre desde la que se pueden desarrollar otros análisis o metodologías dentro y fuera de la geología.

Por temas de conveniencia, extensión y duración, el algoritmo DBSCAN fue el elegido para la realización del análisis de este trabajo. Es posible, sin embargo, que no sea el óptimo o que su ejecución pueda optimizarse con el tiempo. El método desarrollado es más significativo más punto de partida y como enfoque del problema, y menos como "guía metodológica", ya que existen infinidad de algoritmos de distinta complejidad desde los que también se podría abordar el problema dependiendo de las limitaciones computacionales y de tiempo.

Finalmente, teniendo en cuenta los minerales detectados con este trabajo, queda la incógnita de detectar los posibles efectos sobre la salud humana de la exposición a estas partículas en distintas escalas de tiempo, tamaño y concentración. En este trabajo no se discuten dichos efectos por encontrarse en un área del conocimiento que no corresponde al proyecto. Pero en últimas, determinar los efectos sobre la salud de seres vivos (personas, animales y plantas) del material particulado es la pregunta a resolver, y encaminar la investigación en esta dirección, así sea un paso pequeño, es una de las aspiraciones de esta investigación, que por su naturaleza técnica solo puede anteceder aquellas decisiones que impacten la salud pública y el medio ambiente.

8. Conclusiones y recomendaciones

- El algoritmo desarrollado basado en DBSCAN ha demostrado una eficacia para la identificación de muestras que se encuentra en el rango esperado, pudiendo agrupar alrededor de un 20% de las muestras ingresadas.
- En el material particulado muestreado en Bogotá, se identificaron cuatro minerales principales: Calcita, Plagioclasa, Cuarzo y Caolinita, que pueden tener origen geogénico y/o antropogénico.
- Para Cali, se detectaron varios minerales, que incluyen: Cuarzo, Plagioclasa, Moscovita, Hornblenda, Arcillas indeterminadas y Calcita. Estos resultados reflejan una influencia de la mineralogía de rocas básicas presente en su zona occidental.
- En el Valle de Aburrá, se encontraron minerales como Actinolita, Plagioclasa, Olivino, Yeso, Calcita, Ortoclasa, Turmalina, Biotita, Moscovita y Óxidos de hierro, así como aluminosilicatos indeterminados y Antigorita. Esta alta diversidad es esperable ya que en esta zona se realizó un muestreo considerablemente más grande. La mineralogía es acorde a lo reportado para los depósitos aluviales, cuerpos plutónicos y rocas ultramáficas serpentinizadas de la zona. Se destaca la importancia de estudiar la antigorita y otras serpentinas por sus posibles efectos nocivos.
- Se encontró una anomalía de carbono en muchos de los clústeres que puede sugerir que el mayor aporte mineral al PM en las zonas de estudio proviene del desgaste de carreteras asfaltadas. Se requieren más estudios que puedan corroborar esta hipótesis.
- Se requieren más muestreos en diferentes zonas, se recomienda la zona industrial sur de Bogotá y las cercanías a Yumbo – Valle del Cauca.

9. Bibliografía

Amato, F. E. (2018). Non-exhaust emissions: an urban air quality problem for public health; impact and mitigation measures. Academic Press.

Avellaneda Franco, J. D., Rausch, J., Jaramillo Vogel, D., & Duque-Trujillo, J. F. (2020).Caracterización morfo-química y diferenciación del material particulado PM10-2.5 en el Valle de Aburrá (Medellín, Colombia) (Master dissertation, Universidad EAFIT).

Baird, C., & Cann, M. (2012). Environmental chemistry. W.H. Freeman.

Berry, T. A., Belluso, E., Vigliaturo, R., Gieré, R., Emmett, E. A., Testa, J. R., ... & Wallis, S. L. (2022). Asbestos and other hazardous fibrous minerals: potential exposure pathways and associated health risks. International Journal of Environmental Research and Public Health, 19(7), 4031.

Billet, S., Landkocz, Y., Martin, P. J., Verdin, A., Ledoux, F., Lepers, C., André, V., Cazier, F., Sichel, F., Shirali, P., Gosset, P., & Courcot, D. (2018). Chemical characterization of fine and ultrafine PM, direct and indirect genotoxicity of PM and their organic extracts on pulmonary cells. Journal of Environmental Sciences (China), 71, 168–178. https://doi.org/10.1016/j.jes.2018.04.022

Carvajal, J. H., & Navas, O. (2016). Bogotá "Savanna". Landscapes and Landforms of Colombia, 115-126.

Dana, J. D. (1864). Manual of Mineralogy: Including Observations on Mines, Rocks, Reduction of Ores, and the Applications of the Science to the Arts, with 260 Illustrations. Designed for the Use of Schools and Colleges. Peck, White & Peck.

Departamento administrativo nacional de estadística (DANE). (2023). Proyecciones de población - Indicadores demográficos - marzo 2023. Tomado de: https://www.dane.gov.co/files/censo2018/proyecciones-de-poblacion/presentacion-Proypoblacion-IndDemograficos-ActPostCOVID.pdf

DiFrancesco, P. M., Bonneau, D., & Hutchinson, D. J. (2020). The implications of M3C2 projection diameter on 3D semi-automated rockfall extraction from sequential terrestrial laser scanning point clouds. Remote Sensing, 12(11), 1885.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, No. 34, pp. 226-231).

Goldstein, J. I., Newbury, D. E., Michael, J. R., Ritchie, N. W., Scott, J. H. J., & Joy, D. C. (2017). Scanning electron microscopy and X-ray microanalysis. springer.

Gómez, M., Dawidowski, L., Posada, E., & Correa, M. (2011). Chemical composition of PM2. 5 in three zones of the Aburrá Valley, Medellin, Colombia. Proceedings of the Air and Waste Management Association's Annual Conference and Exhibition, AWMA, June, 2534–2545.

Gwenzi, W. (2020). Occurrence, behaviour, and human exposure pathways and health risks of toxic geogenic contaminants in serpentinitic ultramafic geological environments (SUGEs): A medical geology perspective. Science of The Total Environment, 700, 134622.

Hermelín, M., & Rendón, D. A. (2007). Medellín. Entorno natural de, 17, 187-210.

INGEOMINAS & DAGMA (2005). estudio de microzonificación sísmica de Santiago de Cali – Informe No.2-1 Investigaciones Geológicas y Geomorfológicas. Recuperado de: https://www.academia.edu/download/44293512/Informe2.1Geologia_1.pdf

Jiang, N., Yin, S., Guo, Y., Li, J., Kang, P., Zhang, R., & Tang, X. (2018). Characteristics of mass concentration, chemical composition, source apportionment of PM2. 5 and PM10 and health risk assessment in the emerging megacity in China. Atmospheric pollution research, 9(2), 309-321.

Pachauri, T., Singla, V., Satsangi, A., Lakhani, A., & Kumari, K. M. (2013). SEM-EDX characterization of individual coarse particles in Agra, India. Aerosol and Air Quality Research, 13(2), 523-536.

Particle Vision Inc. (2017). Sigma-2 passive sampler for airborne particulate matter. Code of practice. Recuperado de: https://www.particle-vision.ch/images/downloads/Sigma-2_code_of_practice_2017.pdf

Particle Vision Inc. (2020). Faktenblatt zur staubherkunftsbestimmung mittels rem-edx einzelpartikelanalyse auf sigma-2 proben [factsheet]. Recuperado de https://www.particle-vision.ch/images/downloads/Faktenblatt_Sigma-2_Validation_coarse_mode.pdf

Ramírez, O., Sánchez de la Campa, A. M., Amato, F., Catacolí, R. A., Rojas, N. Y., & de la Rosa, J. (2018a). Chemical composition and source apportionment of PM10 at an urban background site in a high–altitude Latin American megacity (Bogota, Colombia). 233, 142–155. https://doi.org/10.1016/j.envpol.2017.10.045

Rausch, J., Jaramillo-Vogel, D., Perseguers, S., Schnidrig, N., Grobéty, B., & Yajan, P. (2022). Automated identification and quantification of tire wear particles (TWP) in airborne dust: SEM/EDX single particle analysis coupled to a machine learning classifier. Science of The Total Environment, 803, 149832.

RMCAB (2021). Informe anual de calidad del aire de Bogotá 2021. PA10-PR04-M2.

SIATA (2021). Informe anual de calidad del aire 2021. Report no. F-GAA-RA-75.

Silva, L. F. O., Milanes, C., Pinto, D., Ramírez, O., & Lima, B. D. (2020). Multiple hazardous elements in nanoparticulate matter from a Caribbean industrialized atmosphere. Chemosphere, 239, 124776. https://doi.org/10.1016/j.chemosphere.2019.124776

Vargas, F. A., Rojas, N. Y., Pachon, J. E., & Russell, A. G. (2012). PM10 characterization and source apportionment at two residential areas in Bogota. Atmospheric Pollution Research, 3(1), 72–80. https://doi.org/10.5094/APR.2012.006

Velásquez, A., & Prieto, A. (2007). Cali. Entorno natural de 17 ciudades de Colombia, 1, 117.

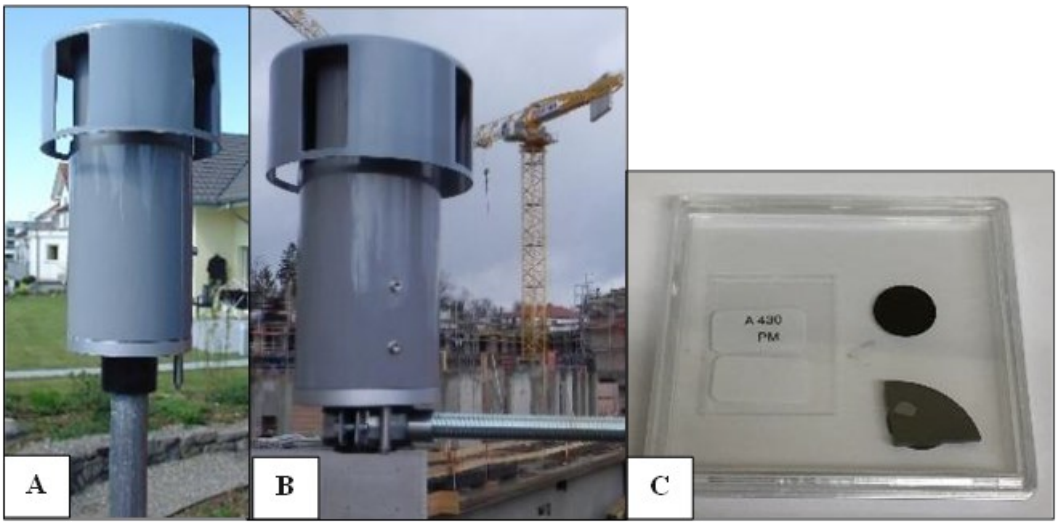
Wenner, D. B., & Taylor Jr, H. P. (1974). D/H and O18/O16 studies of serpentinization of ultramaflc rocks. Geochimica et Cosmochimica Acta, 38(8), 1255-1286.

WHO, W. H. O. (2016). Ambient air pollution: A global assessment of exposure and burden of disease.

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. IEEE Transactions on neural networks, 16(3), 645-678.

Anexos

Figuras



Anexo 1. *(A y B) Muestreadores Sigma-2. (C) Sustratos de boro para muestreo. Tomado de Particle Vision inc. (2017).*

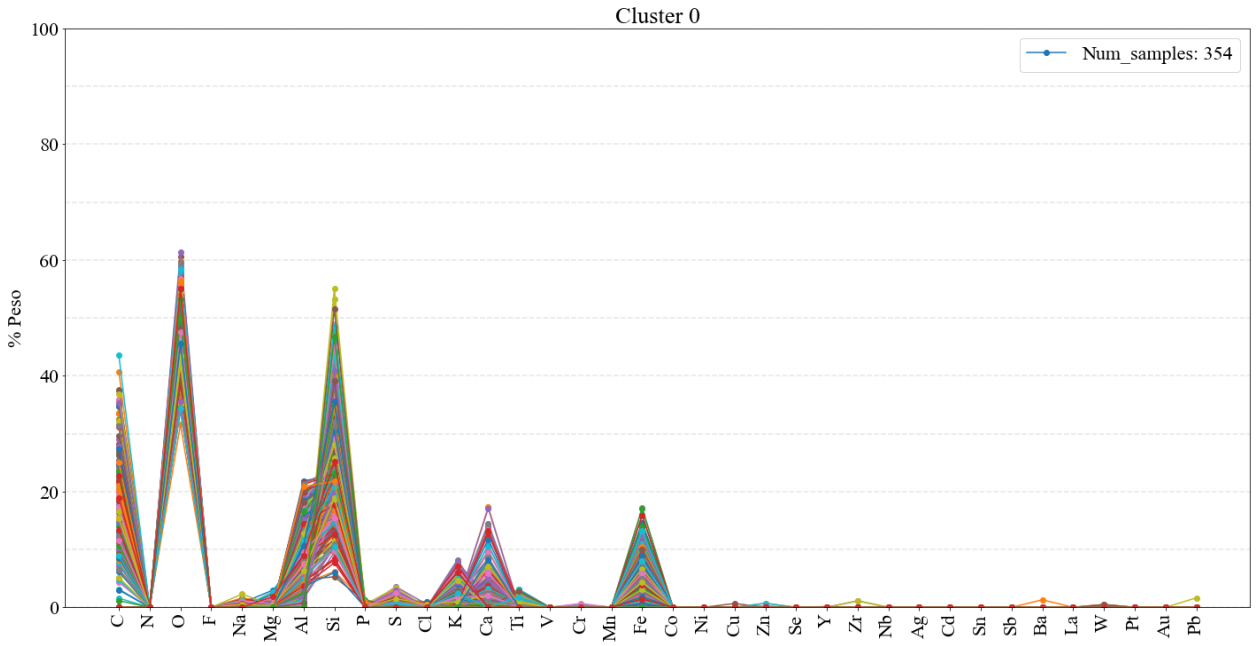
H																	He
Li	Be											B	C	N	O	F	Ne
Na	Mg											Al	Si	P	S	Cl	Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	Lantánidos	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra	Actínidos	Rf	Db	Sg	Bh	Hs	Mt	Ds	Uuu	Cn	Nh	Fl	Mc	Lv	Ts	Og

La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr

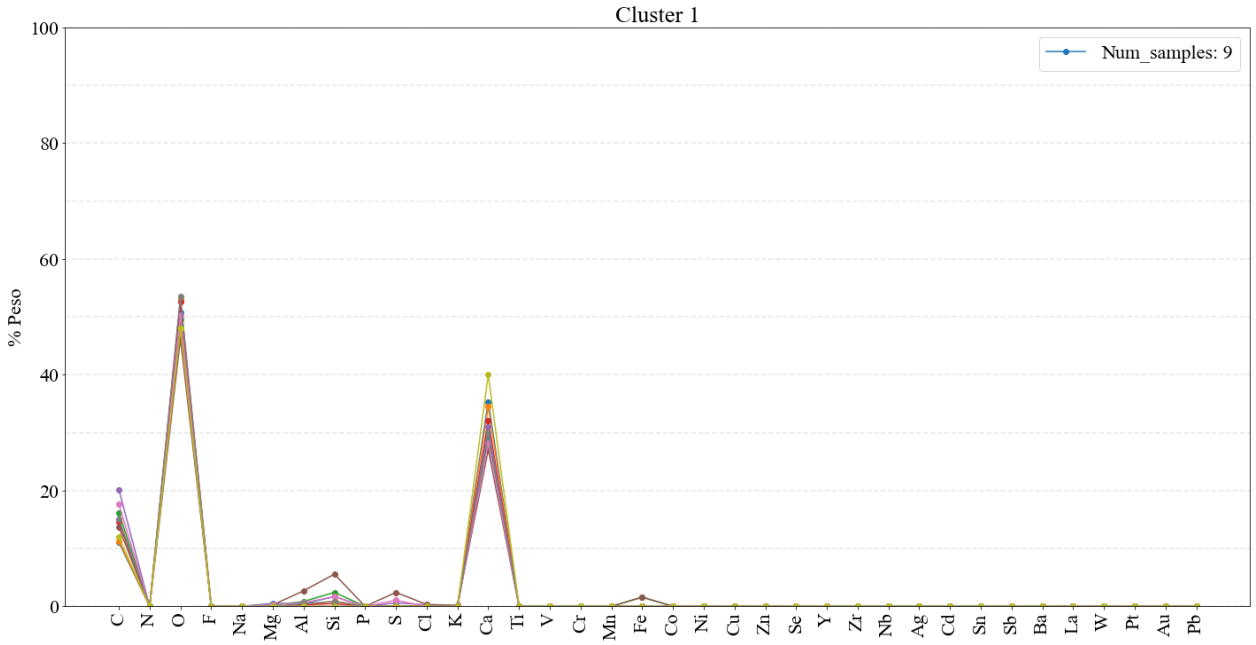
Anexo 2. *Tabla periódica con los elementos eliminados (Negro) y conservados (Color) en la base de datos de partículas minerales. elaboración propia.*

Figuras de resultados del clustering

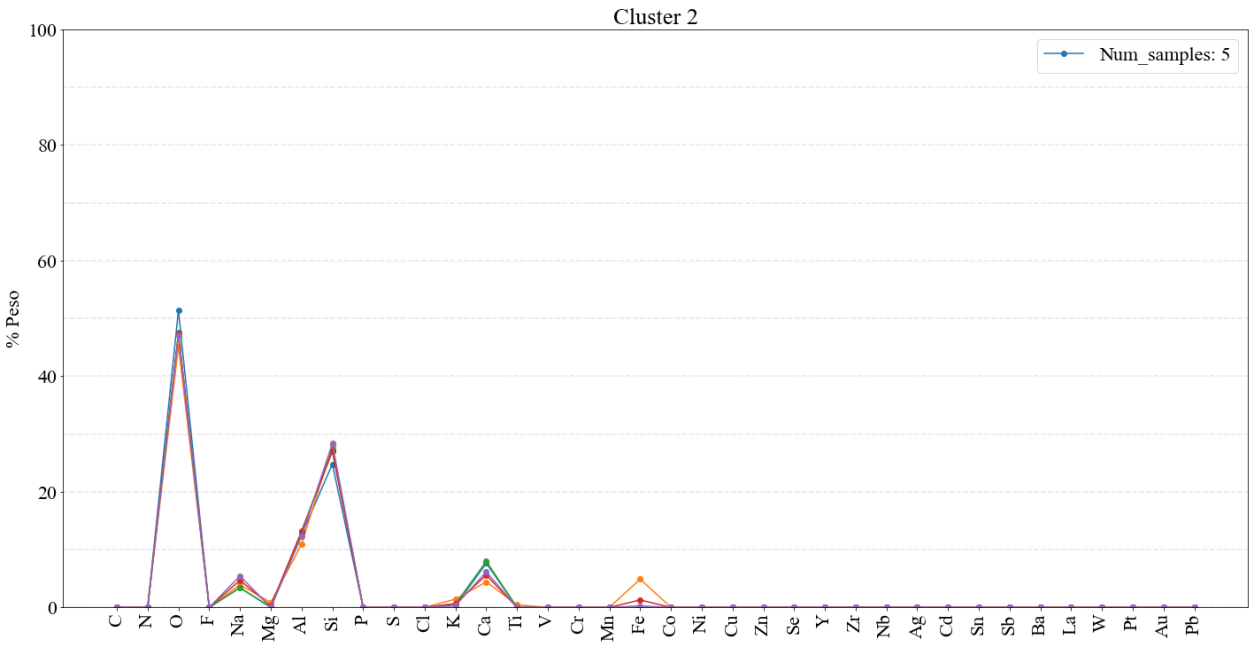
Bogotá



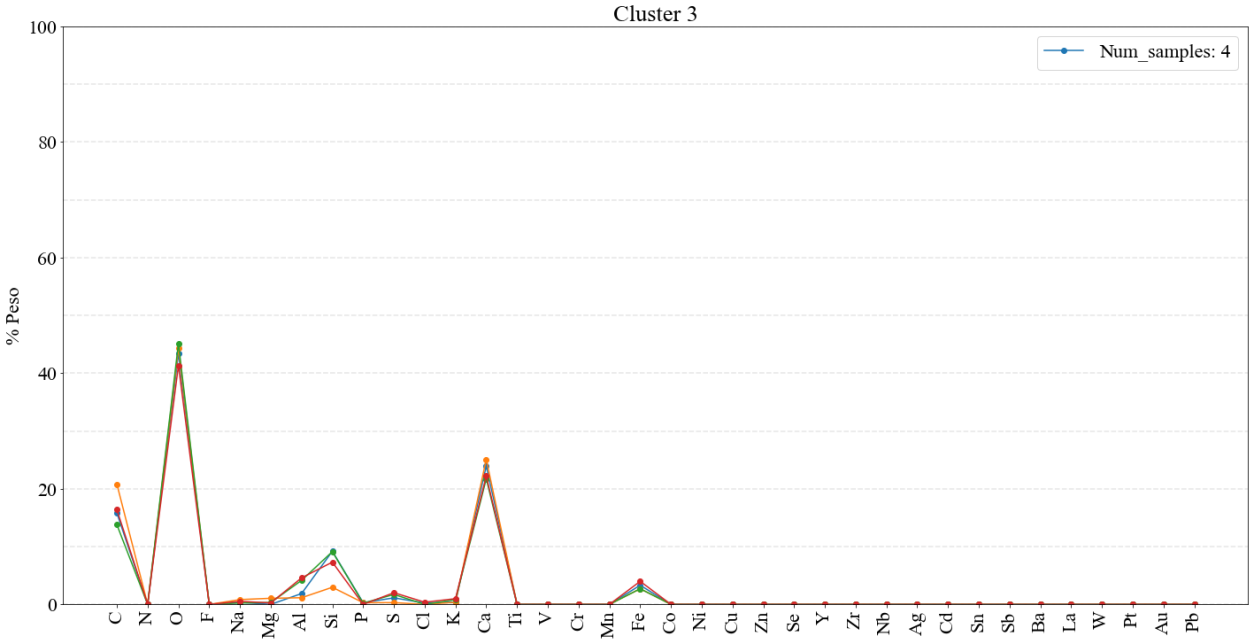
Anexo 3. Bogotá. Iteración 1. Clúster 0. ***(este se tomó para la iteración 2).*** *Elaboración propia.*



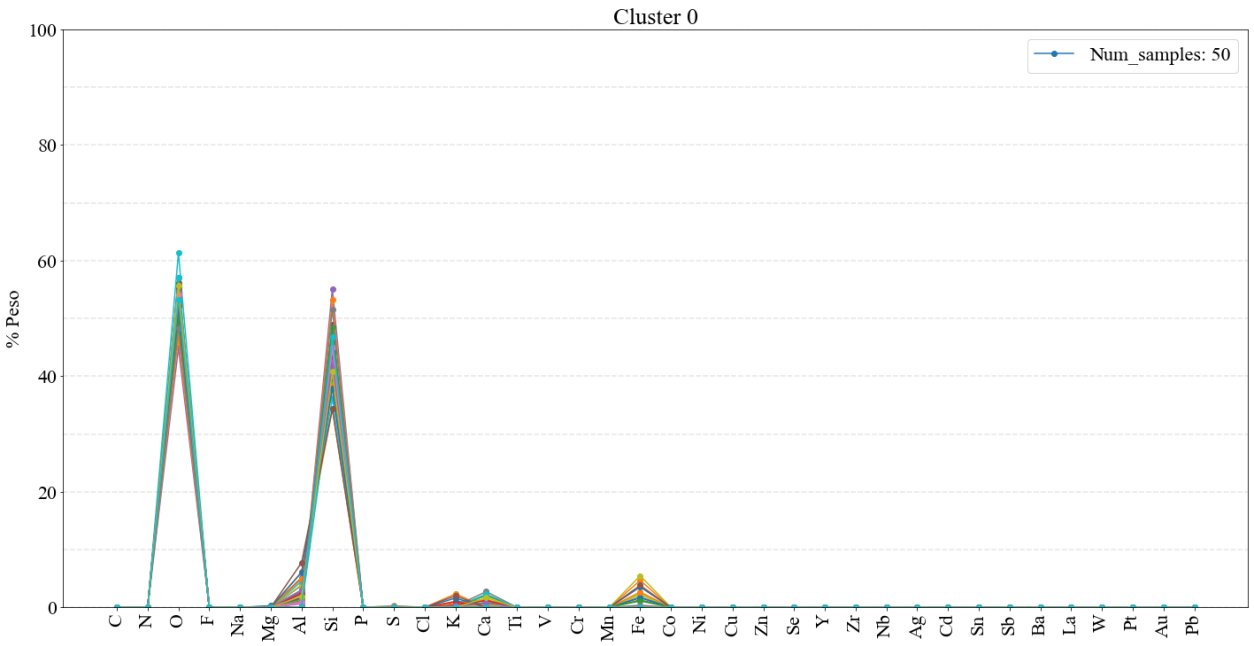
Anexo 4. Bogotá. Iteración 1. Clúster 1. Elaboración propia.



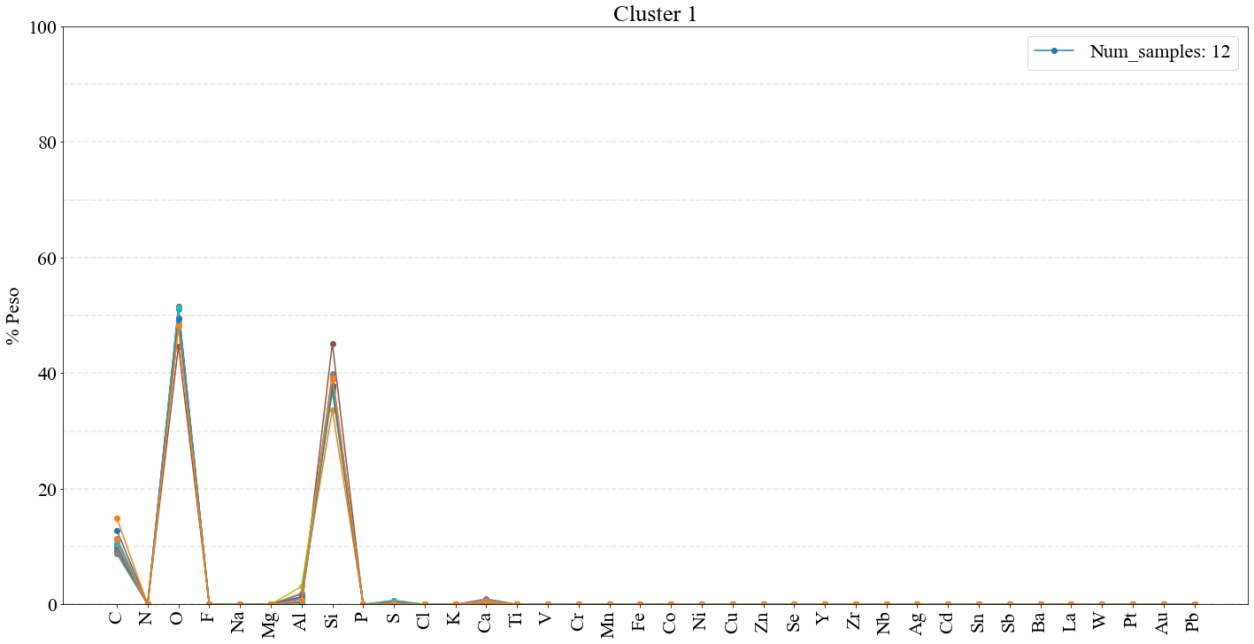
Anexo 5. Bogotá. Iteración 1. Clúster 2. Elaboración propia.



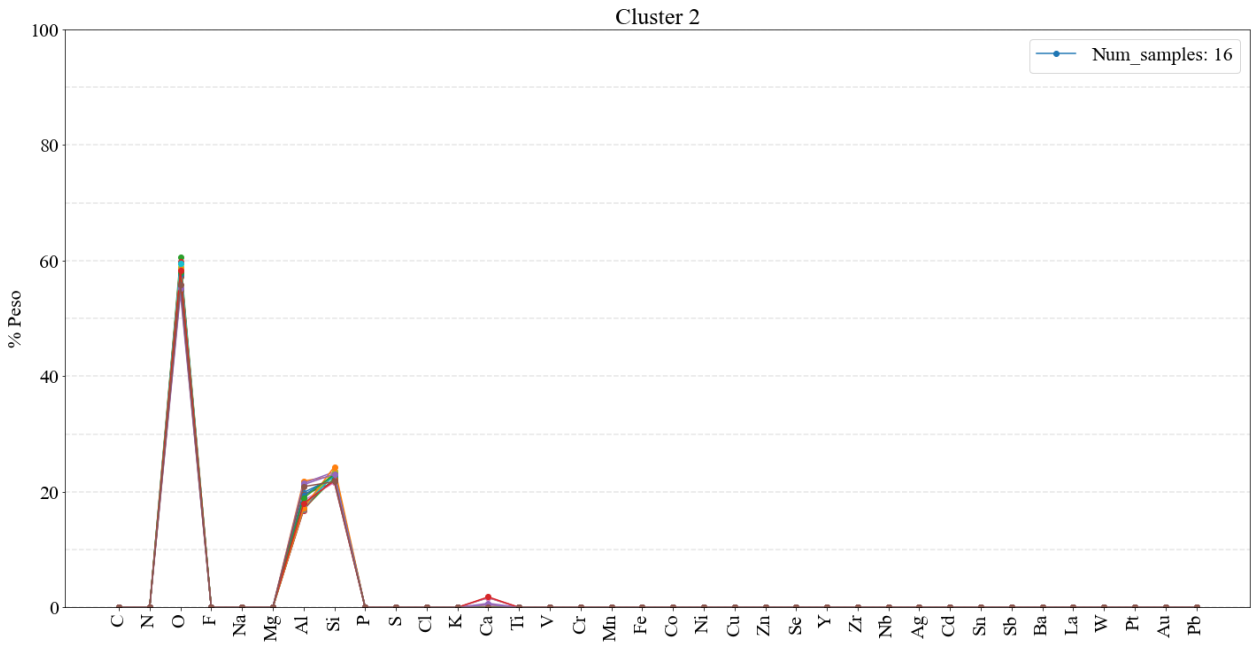
Anexo 6. Bogotá. Iteración 1. Clúster 3. Elaboración propia.



Anexo 7. Bogotá. Iteración 2. Clúster 0. Elaboración propia.

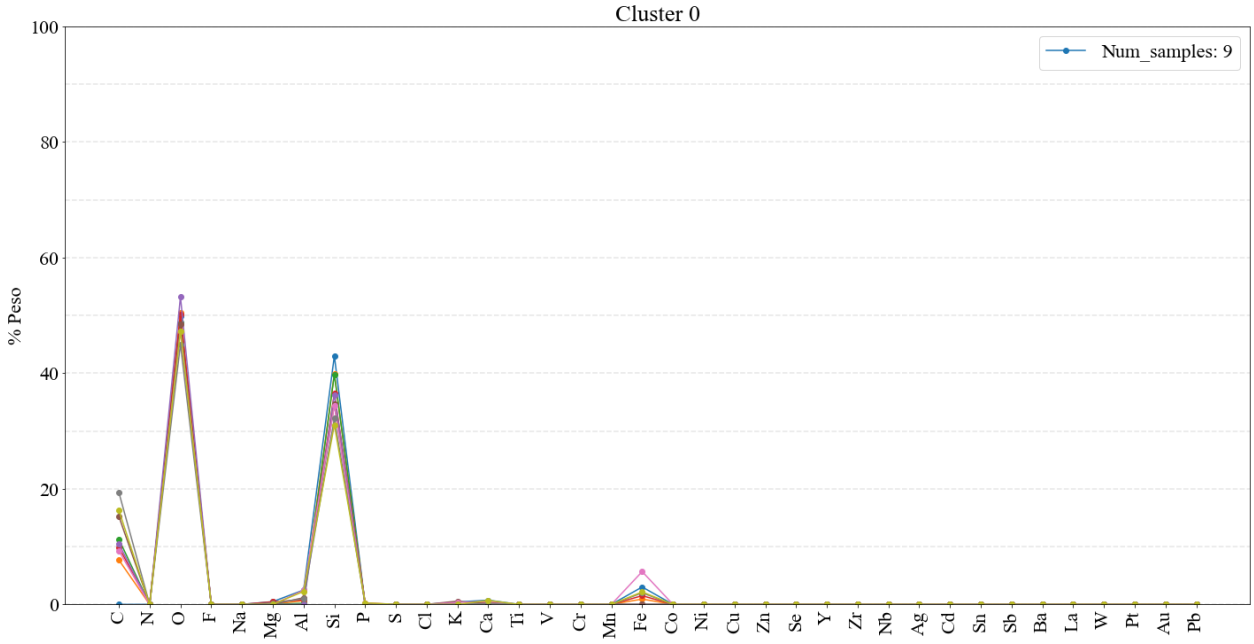


Anexo 8. Bogotá. Iteración 2. Clúster 1. Elaboración propia.

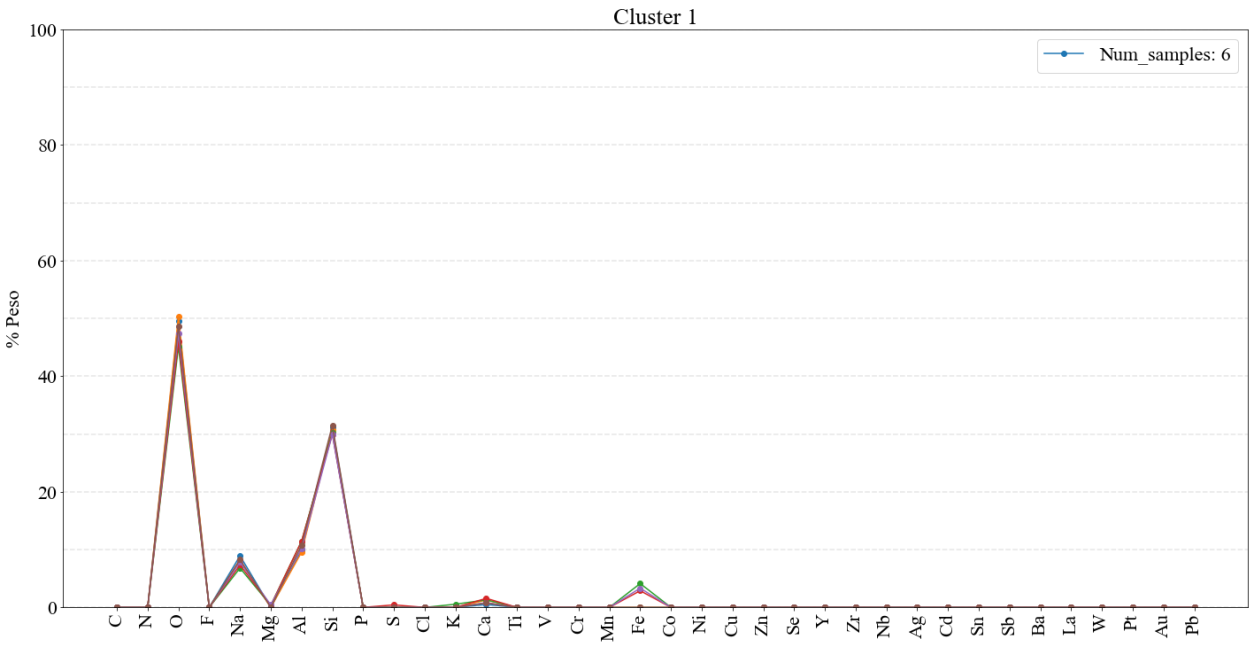


Anexo 9. Bogotá. Iteración 2. Clúster 2. Elaboración propia.

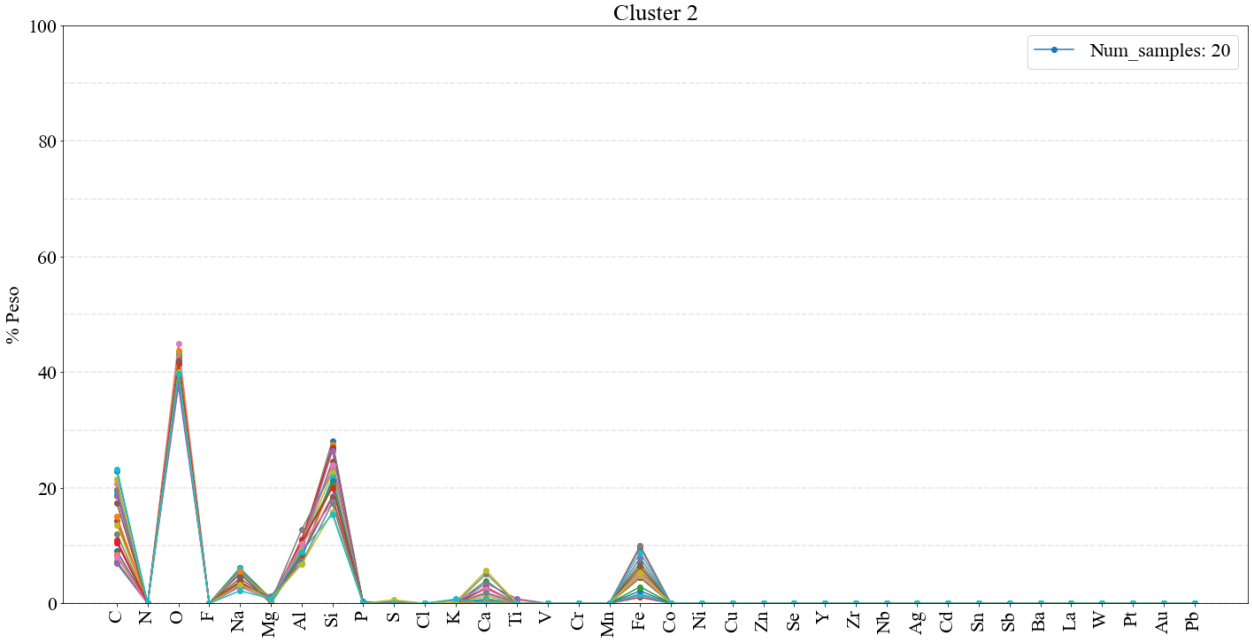
Cali



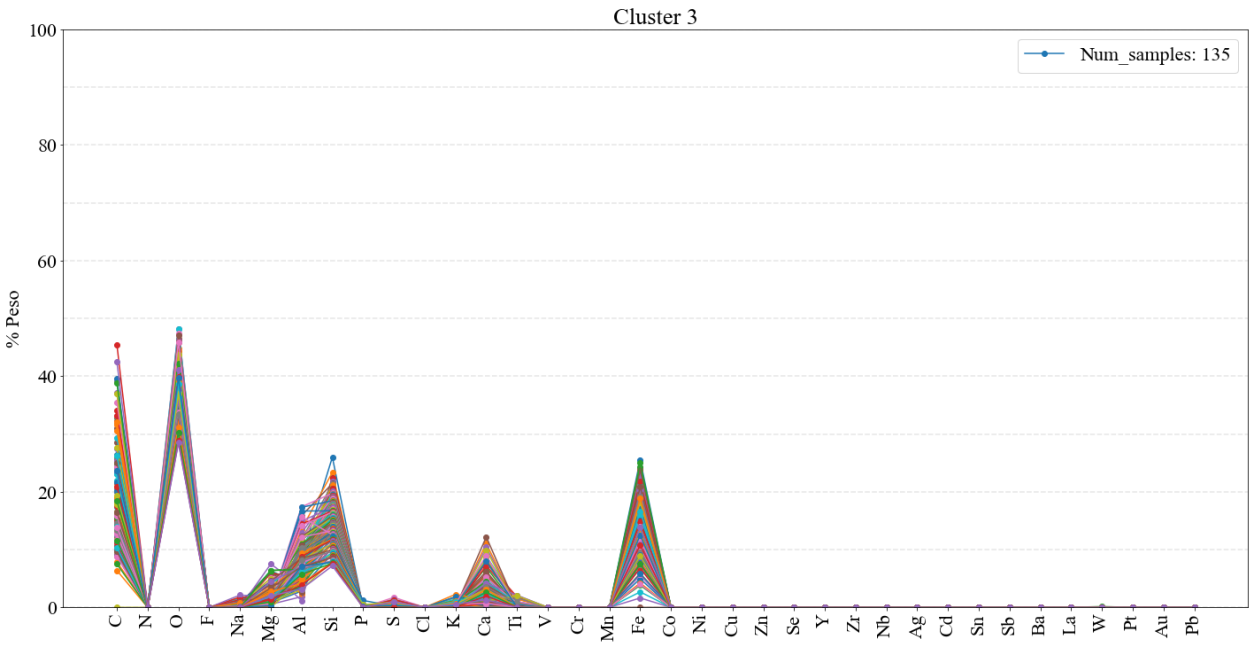
Anexo 10. Cali. Iteración 1. Clúster 0. Elaboración propia.



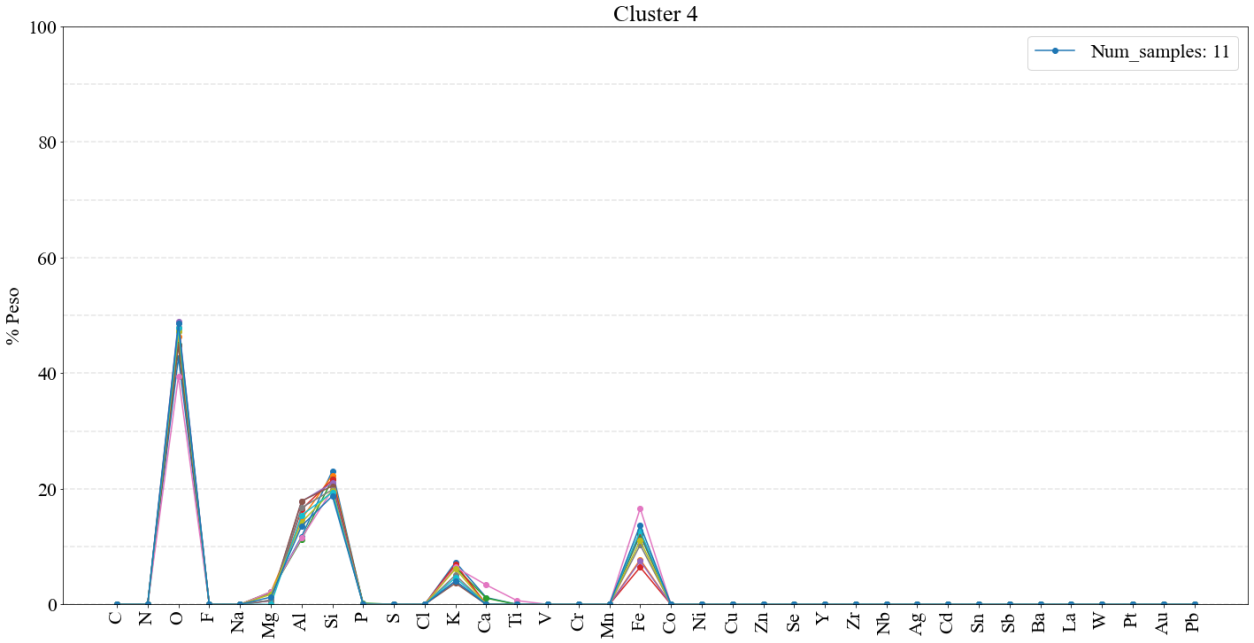
Anexo 11. Cali. Iteración 1. Clúster 1. Elaboración propia.



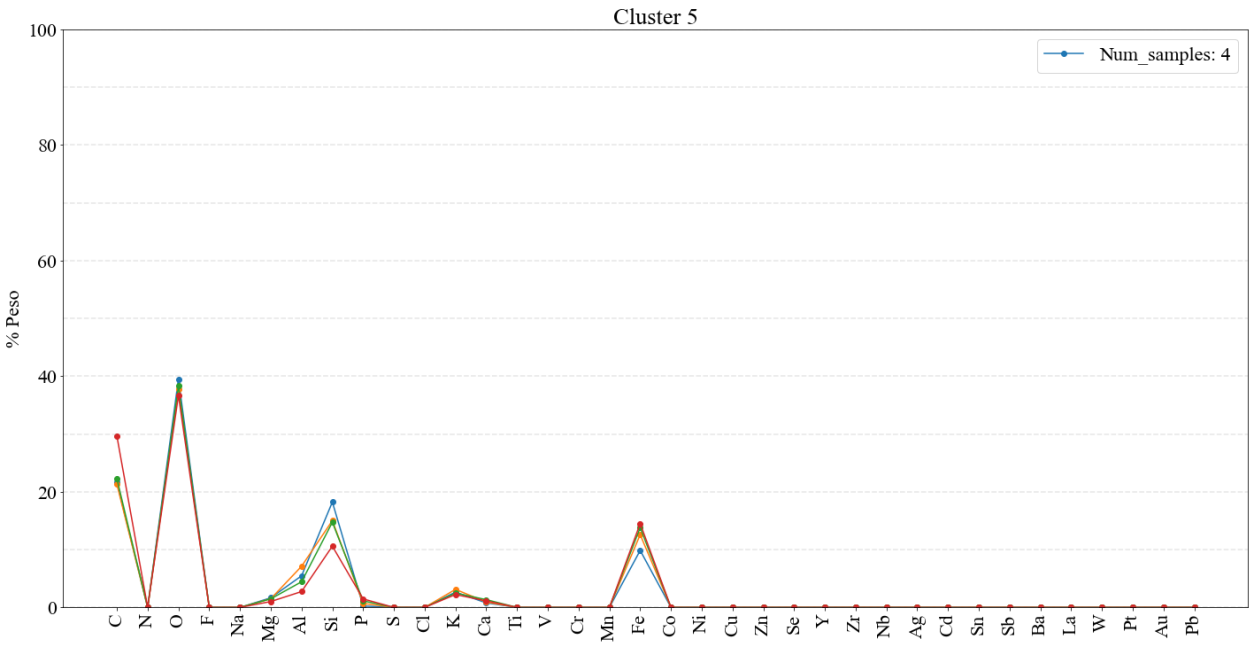
Anexo 12. Cali. Iteración 1. Clúster 2. Elaboración propia.



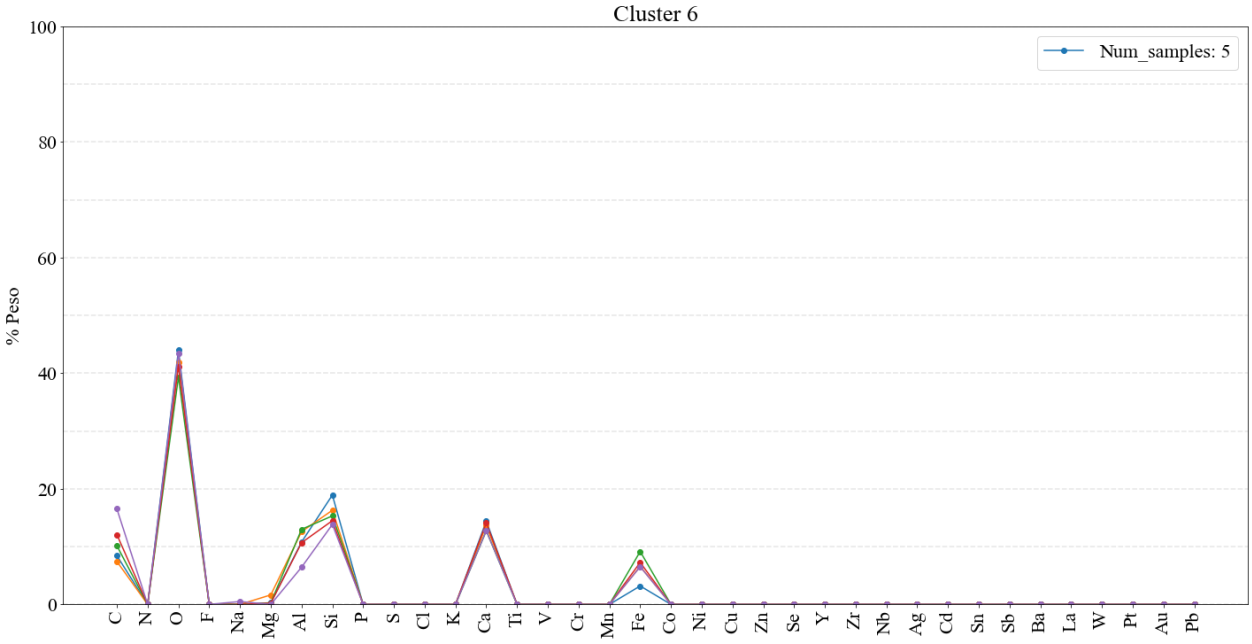
Anexo 13. Cali. Iteración 1. Clúster 3. ***(Este se tomó para la iteración 2****). Elaboración propia.*



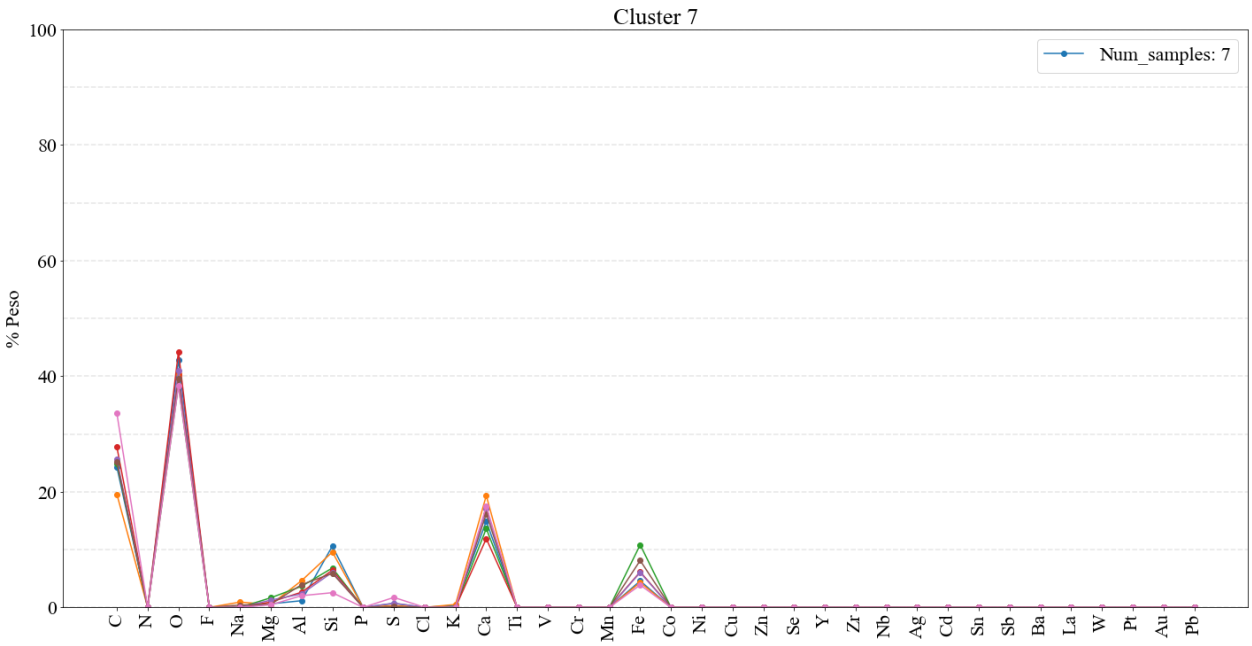
Anexo 14. Cali. Iteración 1. Clúster 4. Elaboración propia.



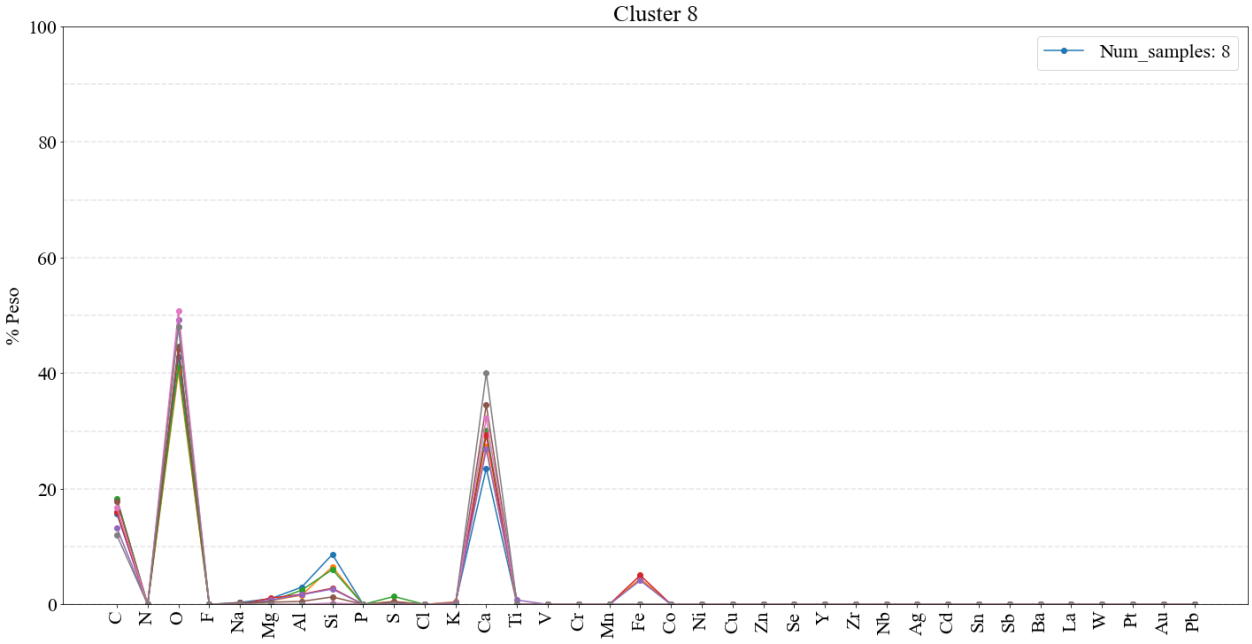
Anexo 15. Cali. Iteración 1. Clúster 5. Elaboración propia.



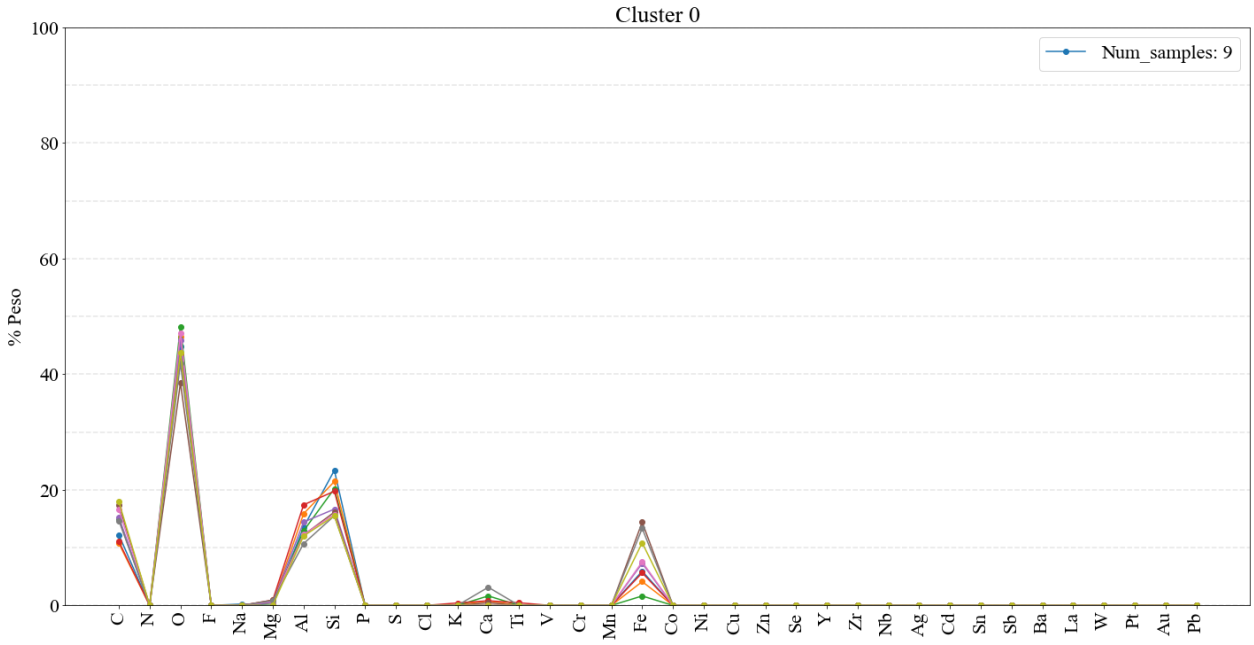
Anexo 16. Cali. Iteración 1. Clúster 6. Elaboración propia.



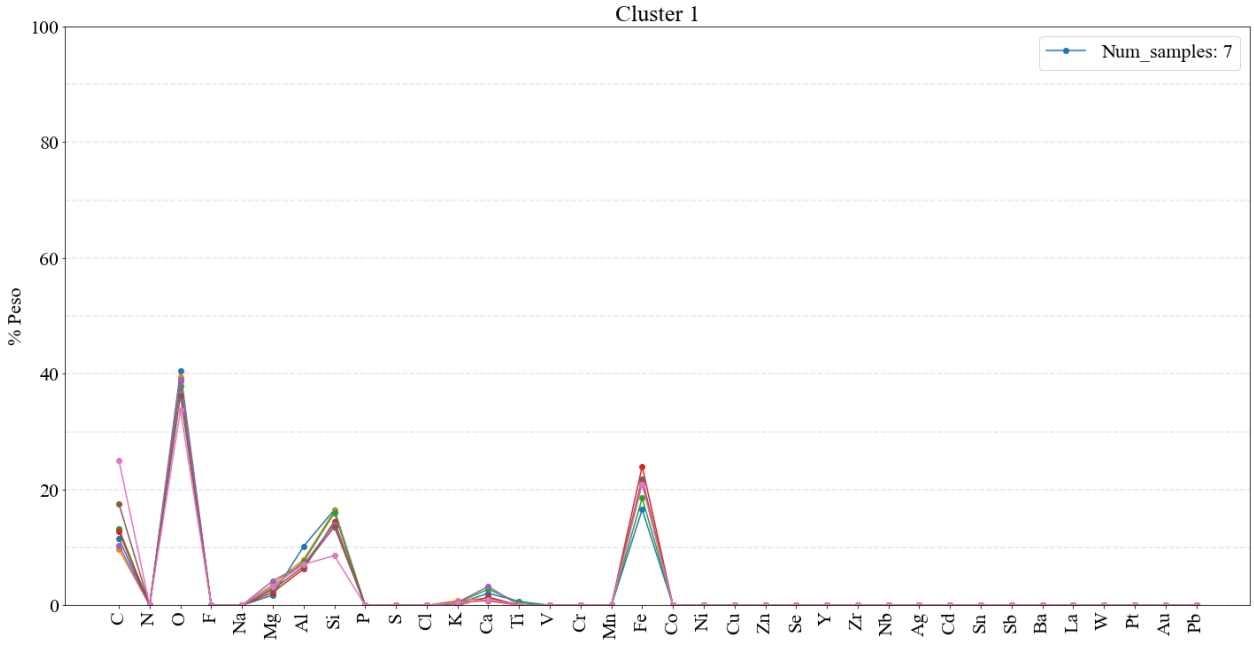
Anexo 17. Cali. Iteración 1. Clúster 7. Elaboración propia.



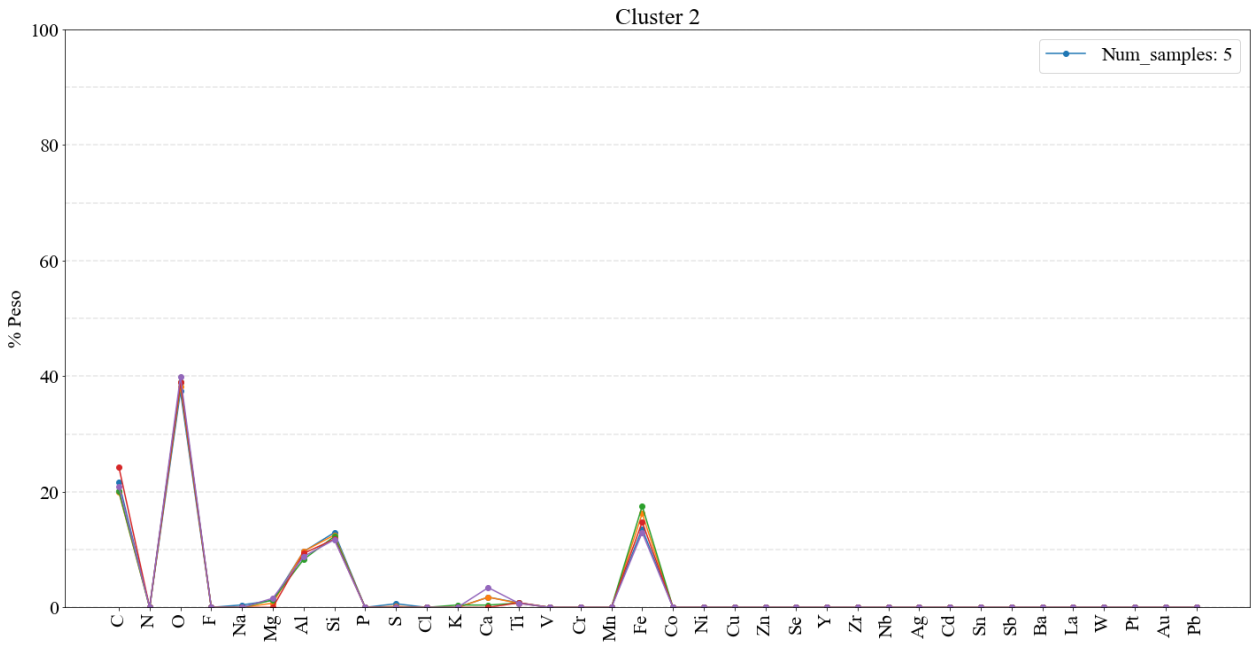
Anexo 18. Cali. Iteración 1. Clúster 8. Elaboración propia.



Anexo 19. Cali. Iteración 2. Clúster 0. Elaboración propia.

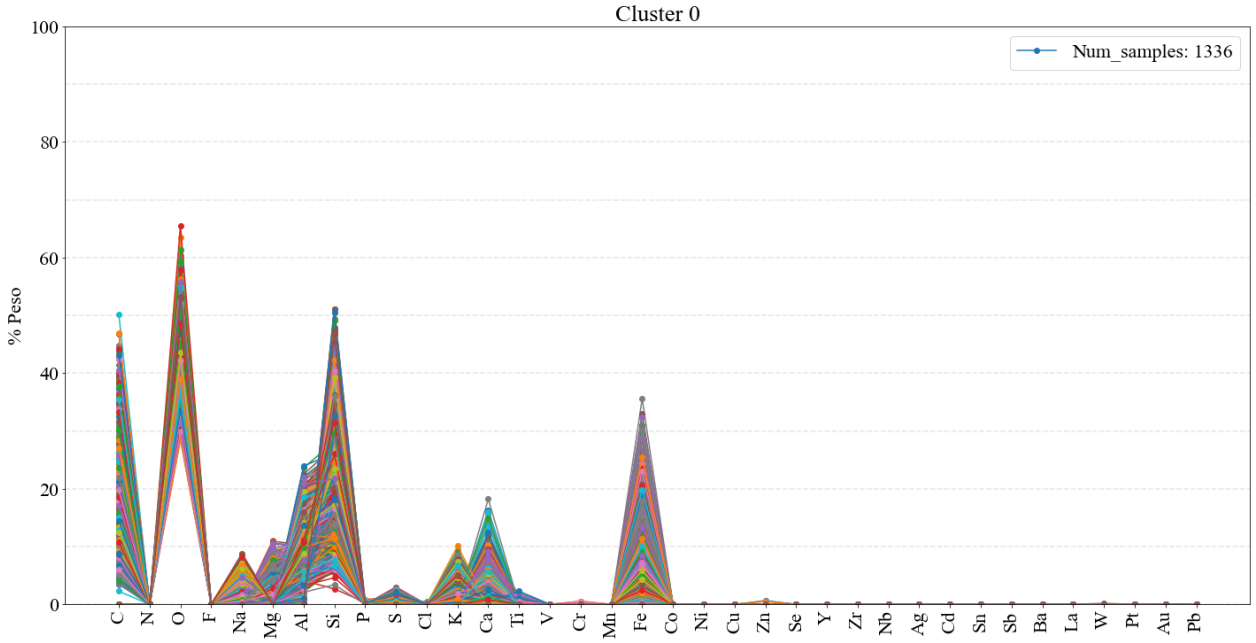


Anexo 20. Cali. Iteración 2. Clúster 1. Elaboración propia.

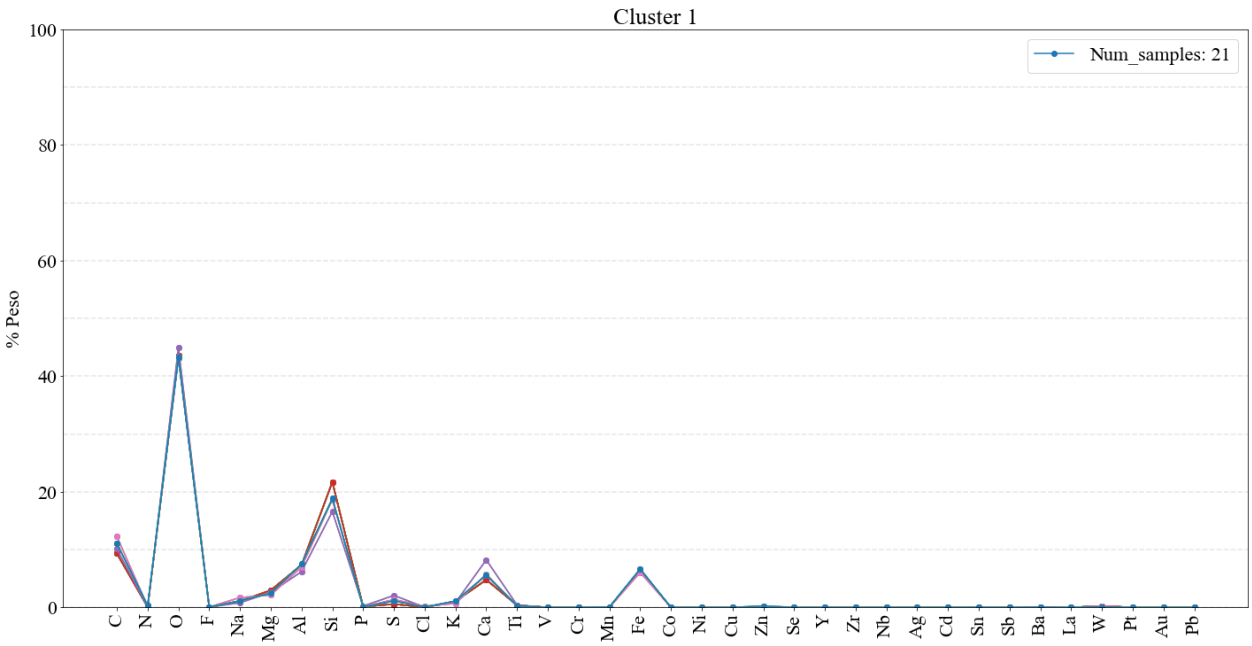


Anexo 21. Cali. Iteración 2. Clúster 2. Elaboración propia.

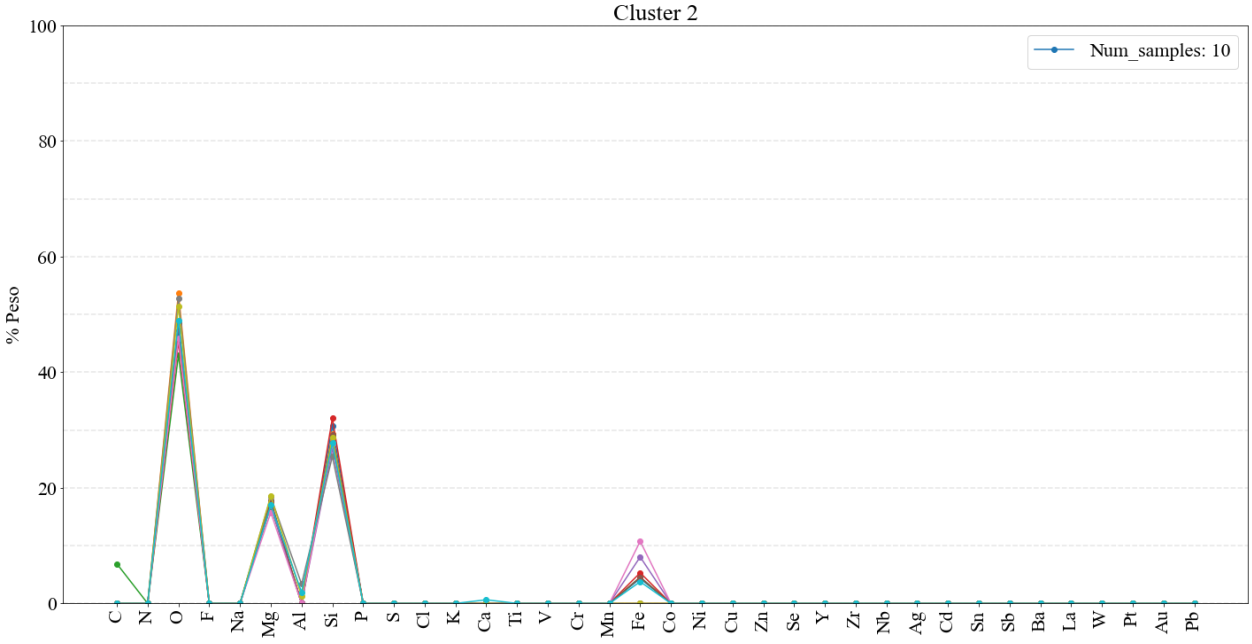
Valle de Aburrá



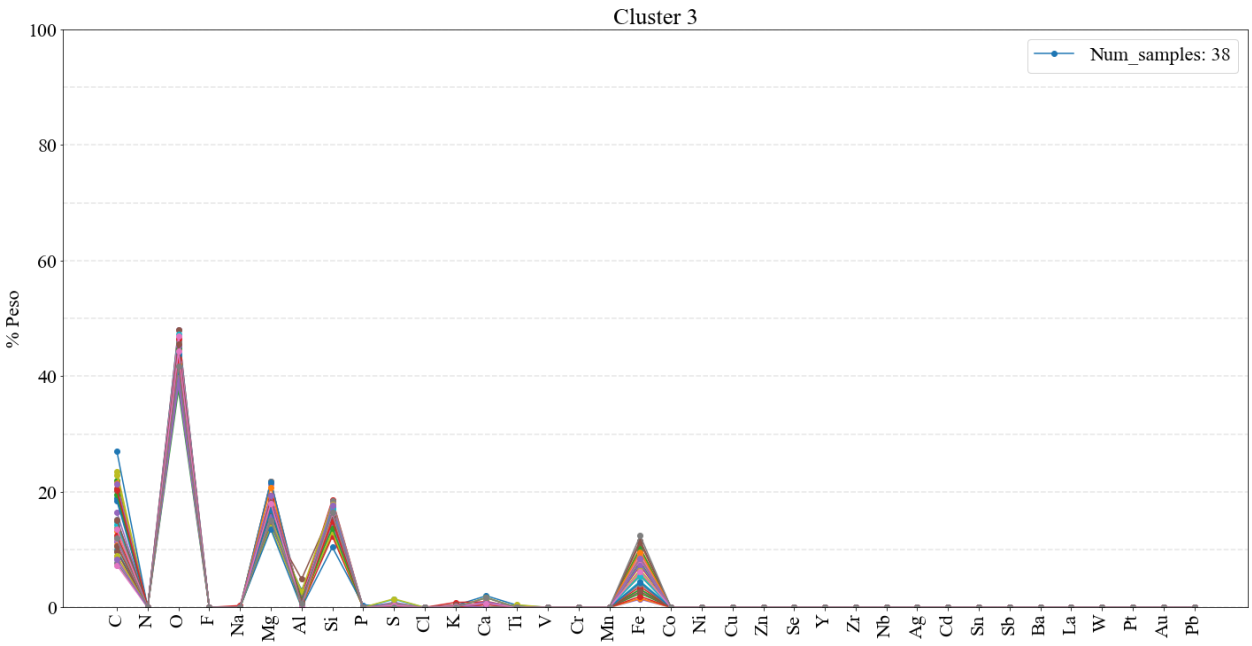
Anexo 22. Valle de Aburrá. Iteración 1. Clúster 0. (este se tomó para la iteración 2). Elaboración propia.



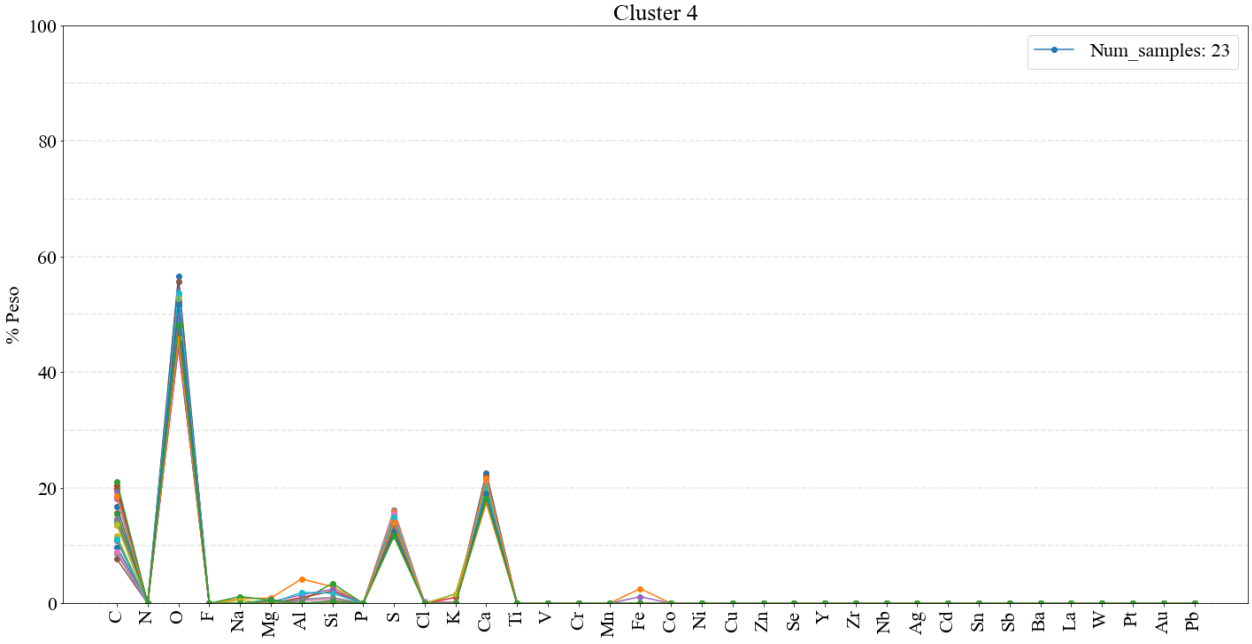
Anexo 23. Valle de Aburrá. Iteración 1. Clúster 1. Elaboración propia.



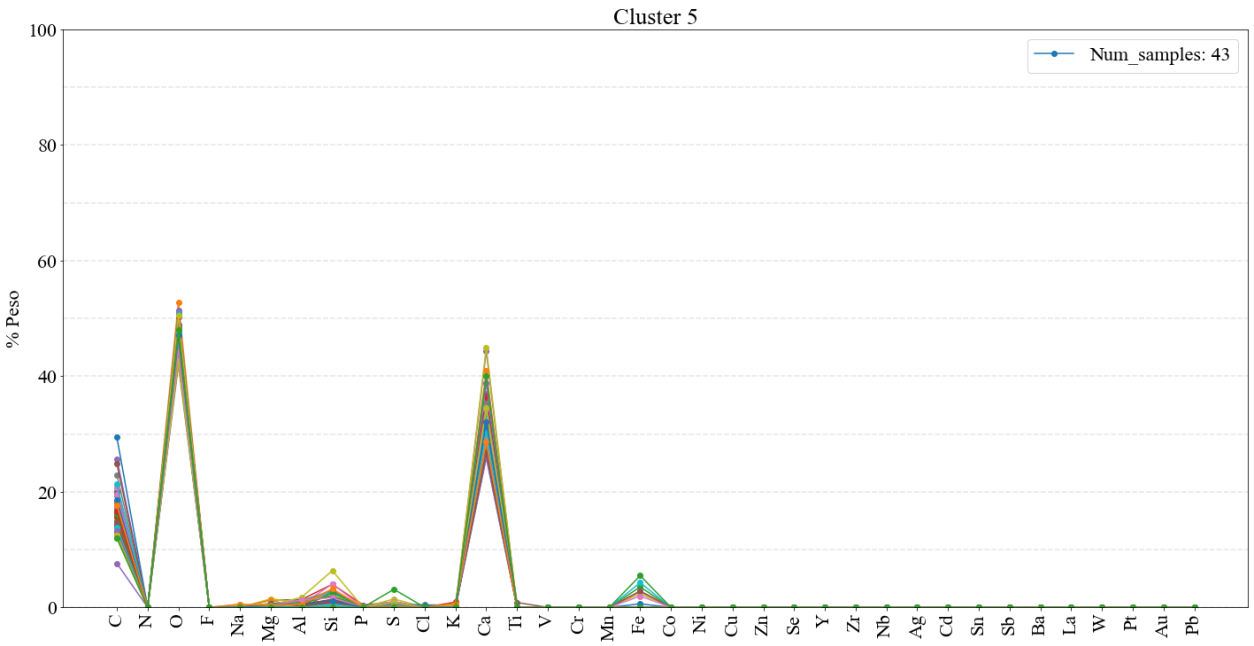
Anexo 24. Valle de Aburrá. Iteración 1. Clúster 2. Elaboración propia.



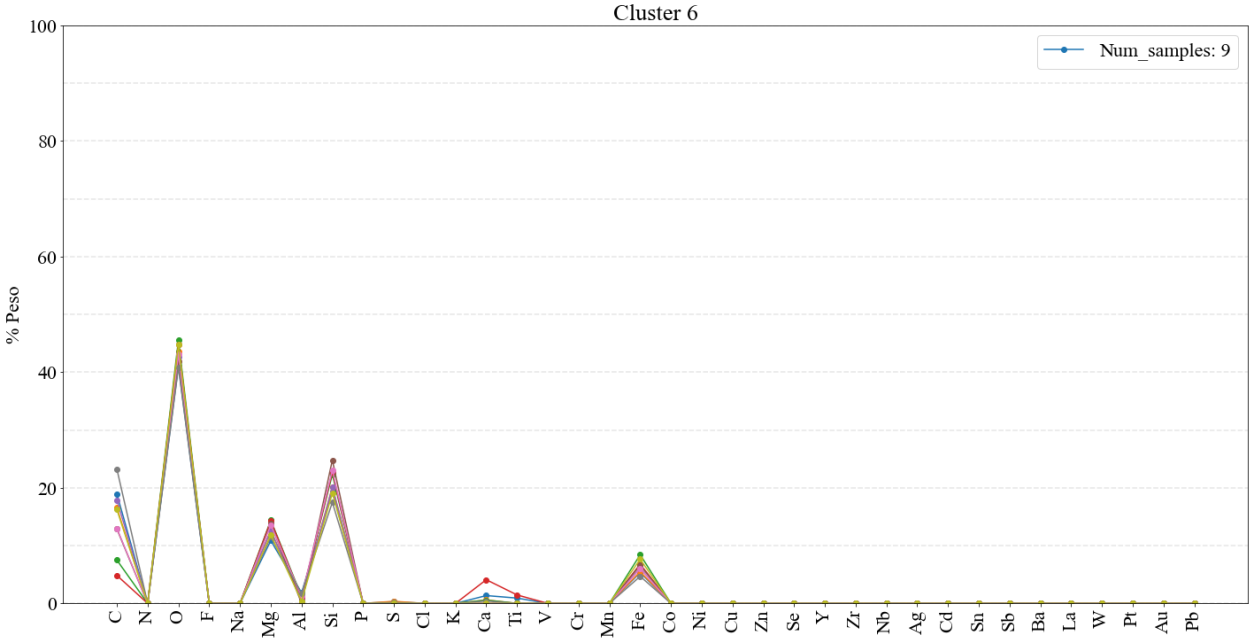
Anexo 25. Valle de Aburrá. Iteración 1. Clúster 3. Elaboración propia.



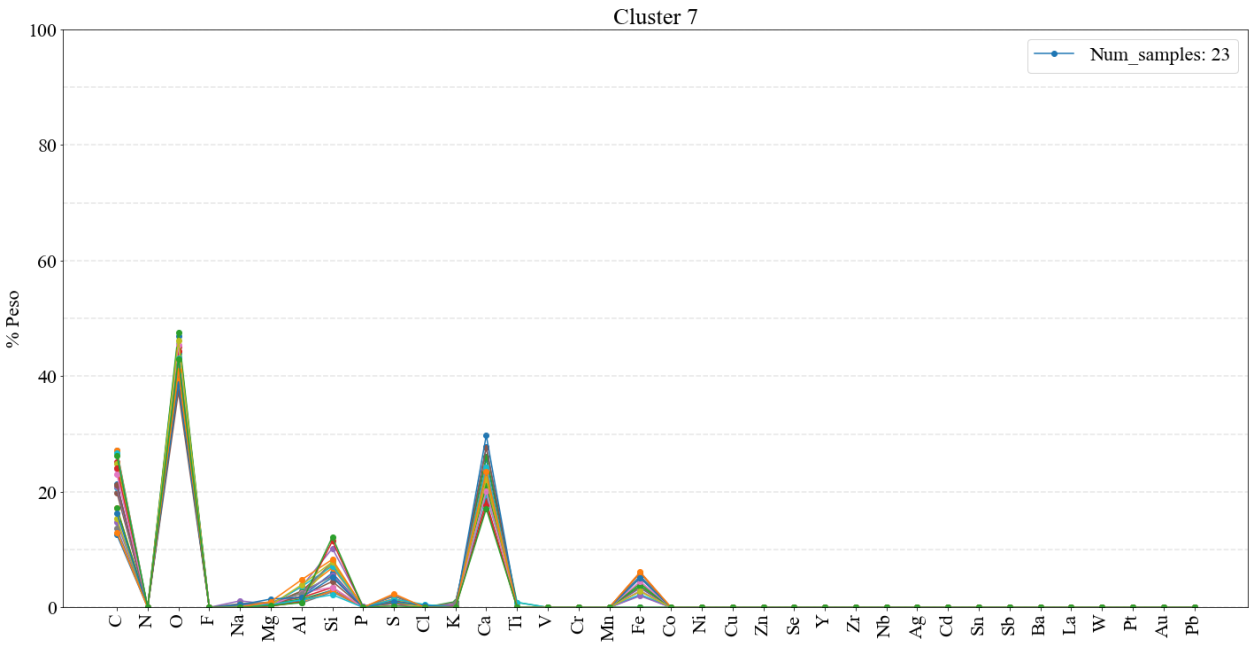
Anexo 26. Valle de Aburrá. Iteración 1. Clúster 4. Elaboración propia.



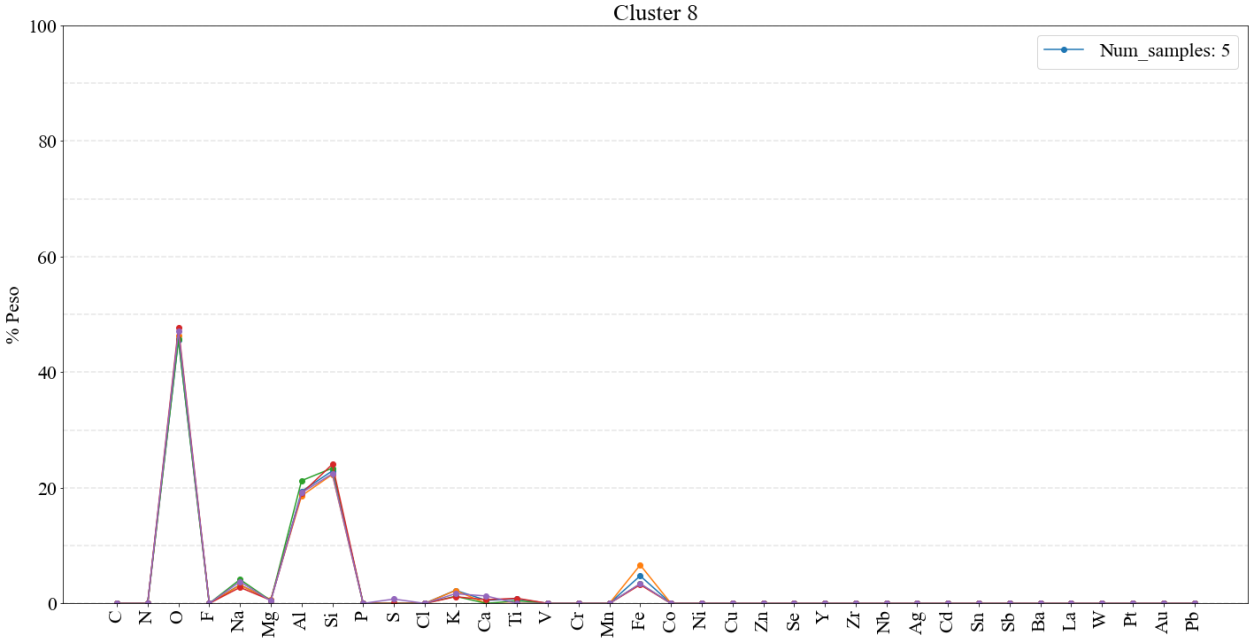
Anexo 27. Valle de Aburrá. Iteración 1. Clúster 5. Elaboración propia.



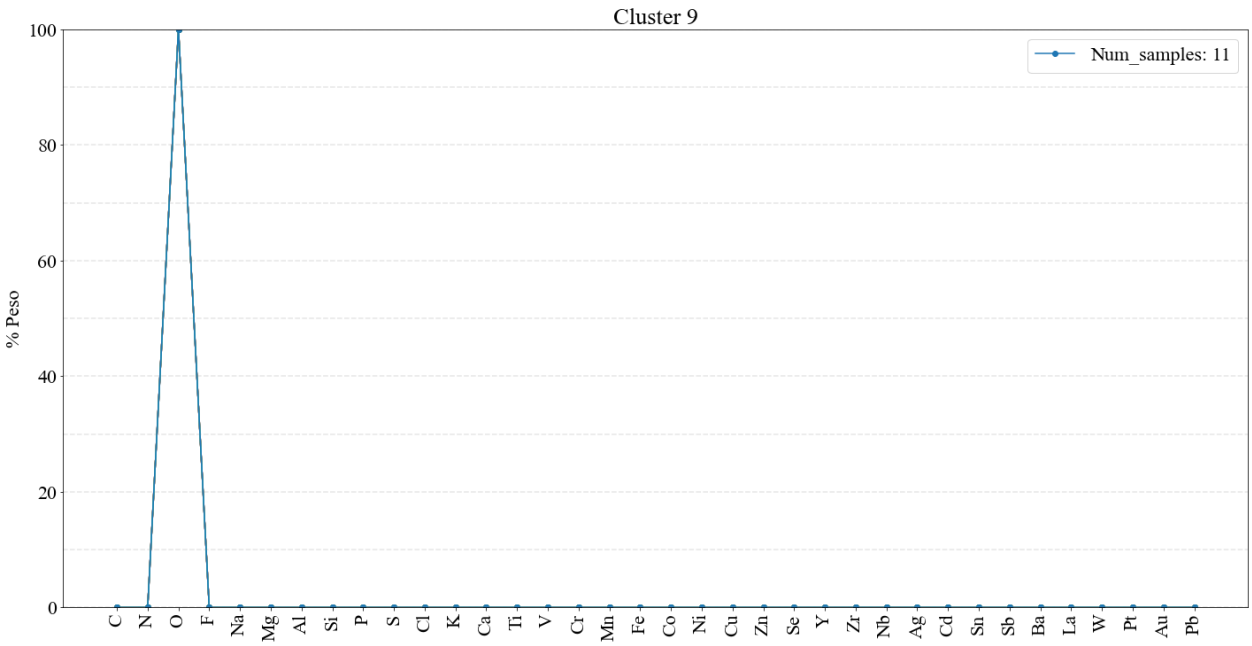
Anexo 28. Valle de Aburrá. Iteración 1. Clúster 6. Elaboración propia.



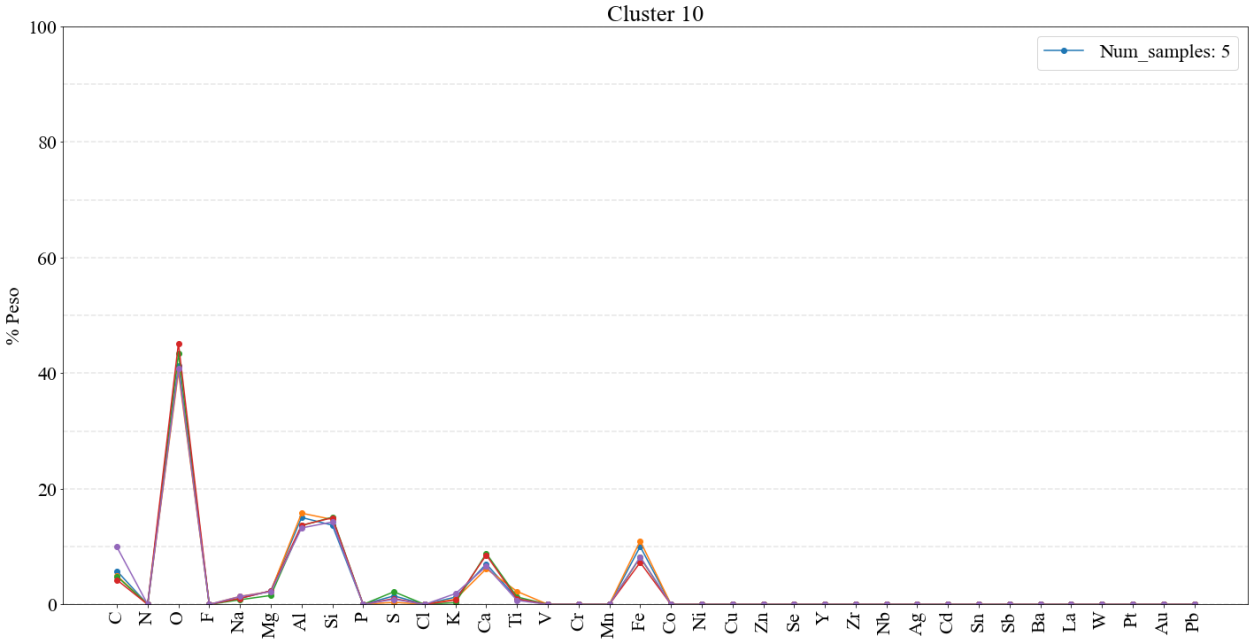
Anexo 29. Valle de Aburrá. Iteración 1. Clúster 7. Elaboración propia.



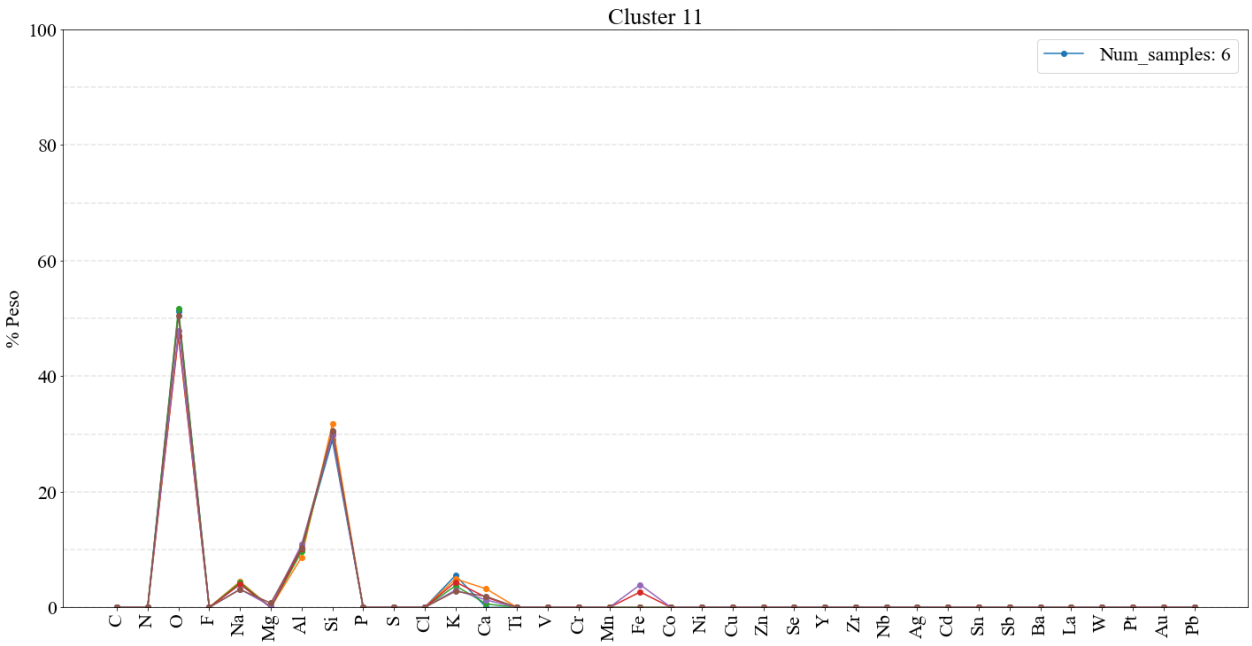
Anexo 30. Valle de Aburrá. Iteración 1. Clúster 8. Elaboración propia.



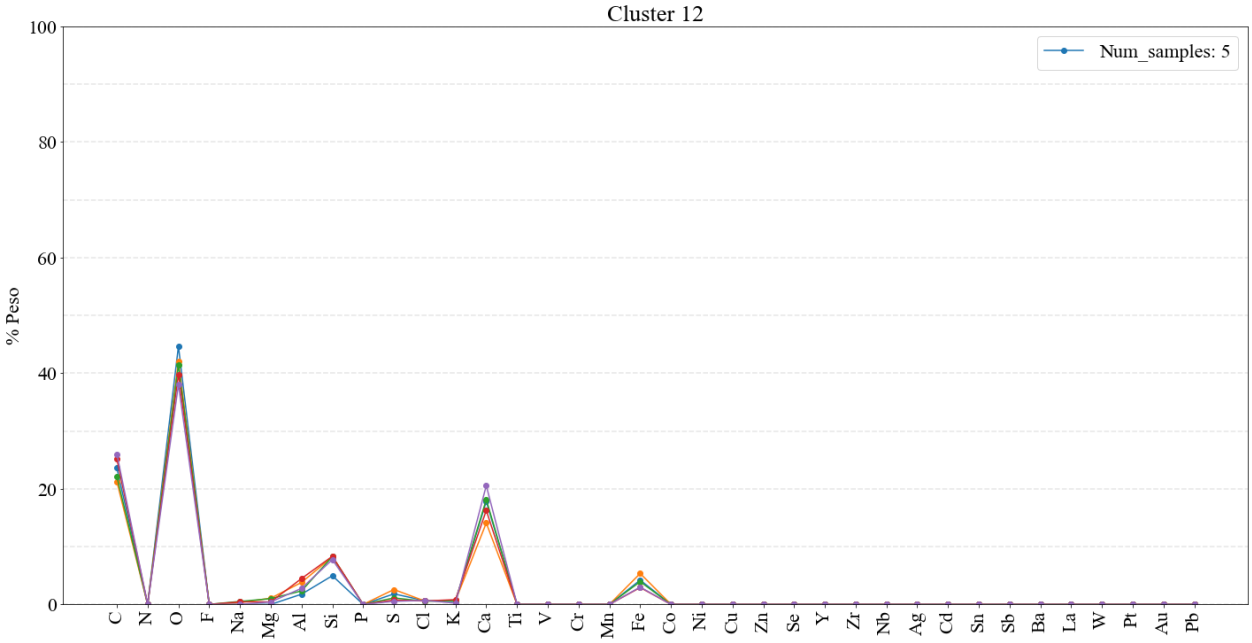
Anexo 31. Valle de Aburrá. Iteración 1. Clúster 9. Elaboración propia.



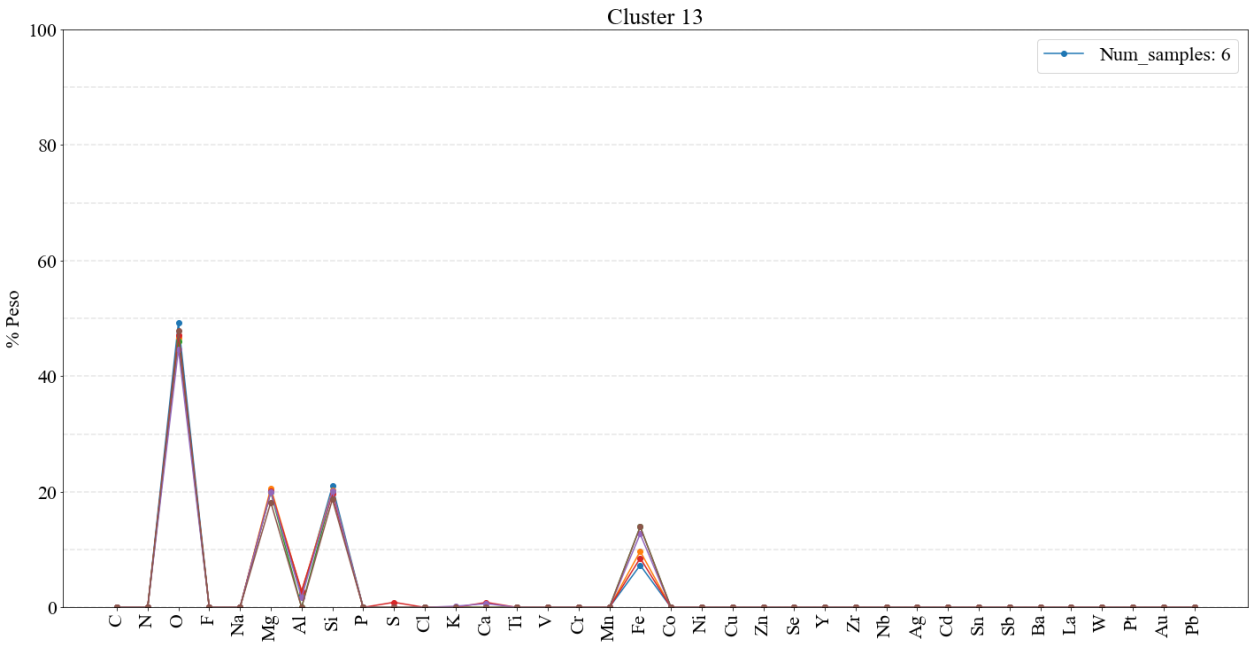
Anexo 32. Valle de Aburrá. Iteración 1. Clúster 10. Elaboración propia.



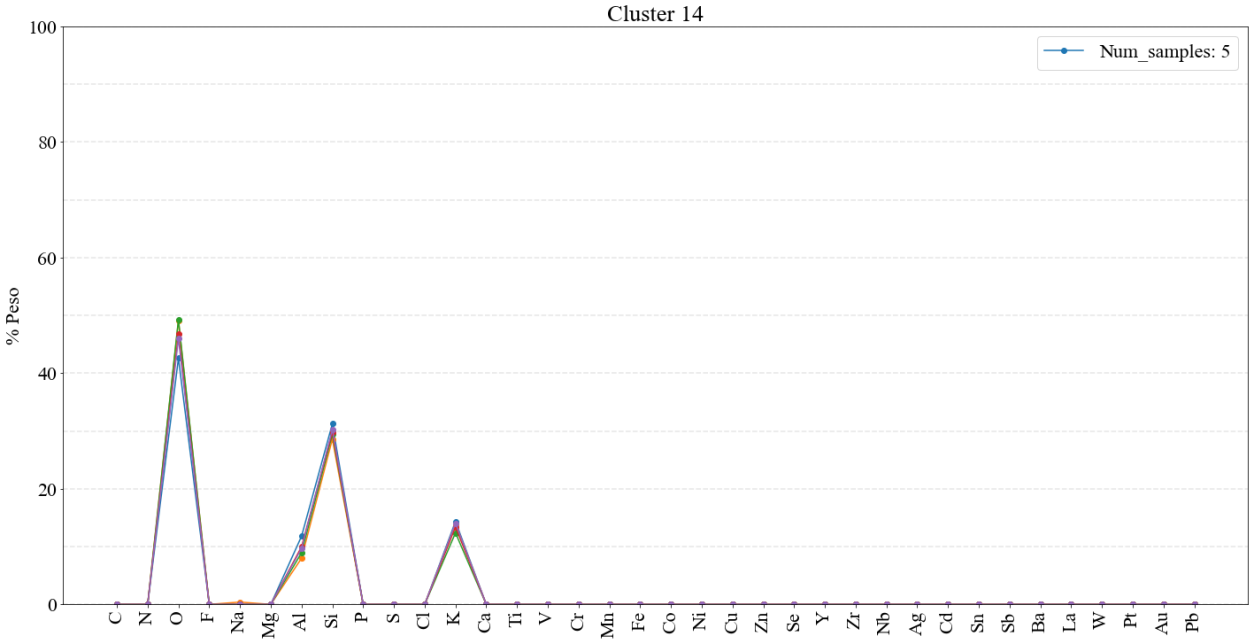
Anexo 33. Valle de Aburrá. Iteración 1. Clúster 11. Elaboración propia.



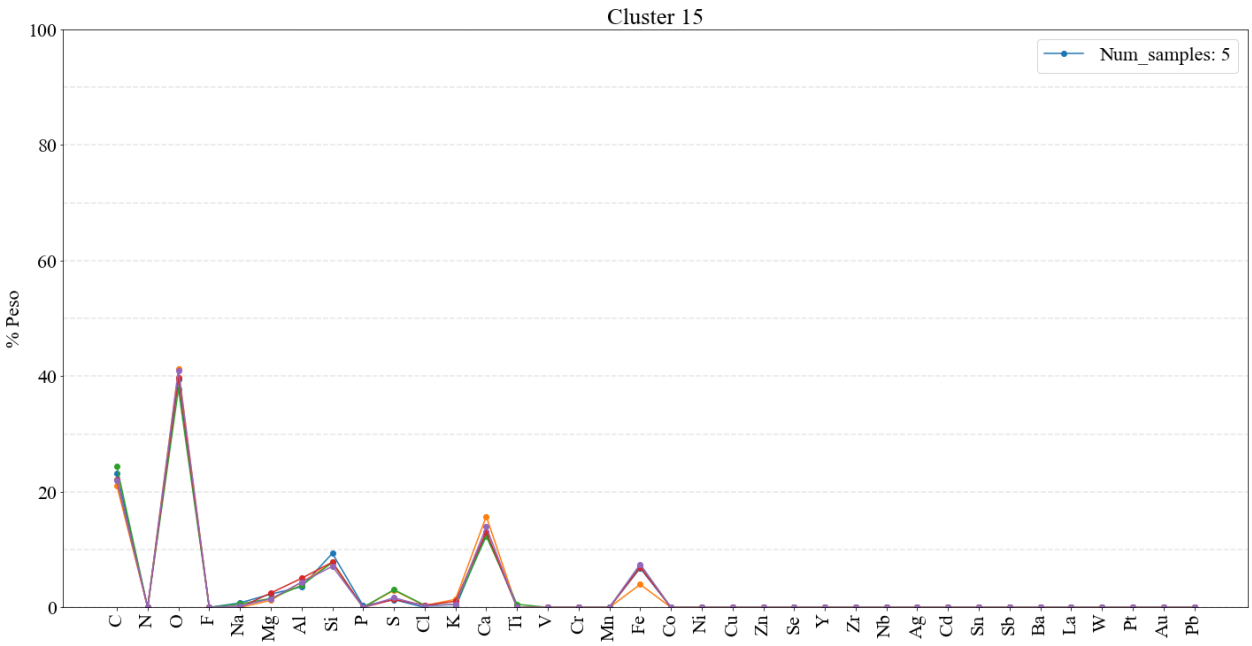
Anexo 34. Valle de Aburrá. Iteración 1. Clúster 12. Elaboración propia.



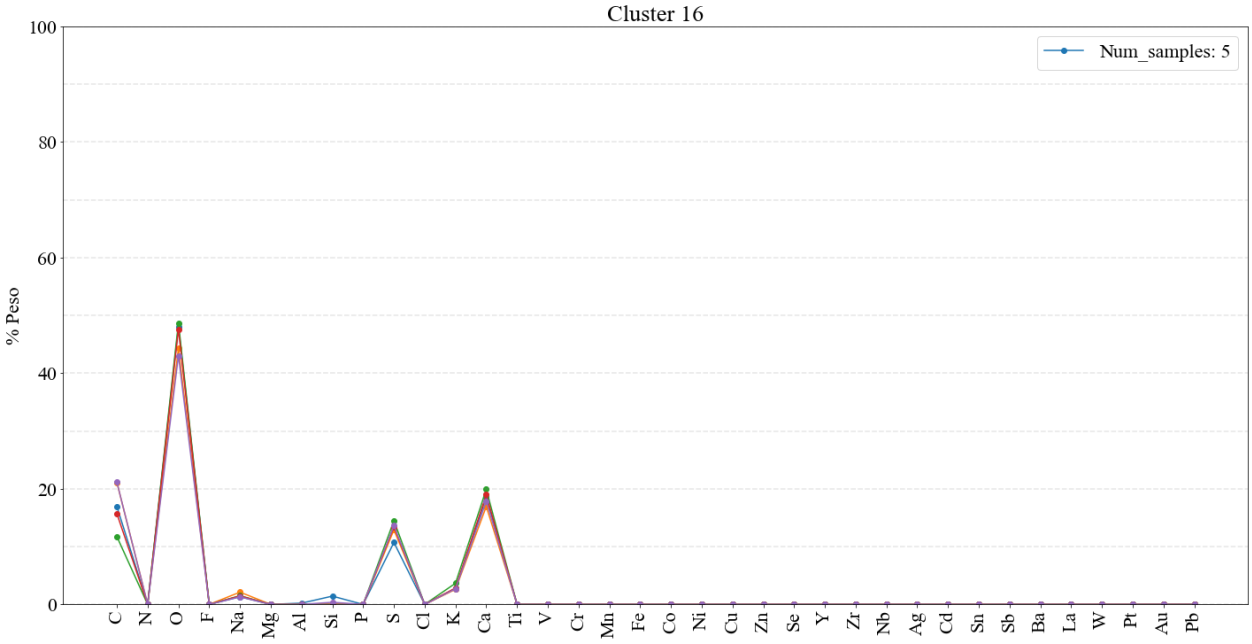
Anexo 35. Valle de Aburrá. Iteración 1. Clúster 13. Elaboración propia.



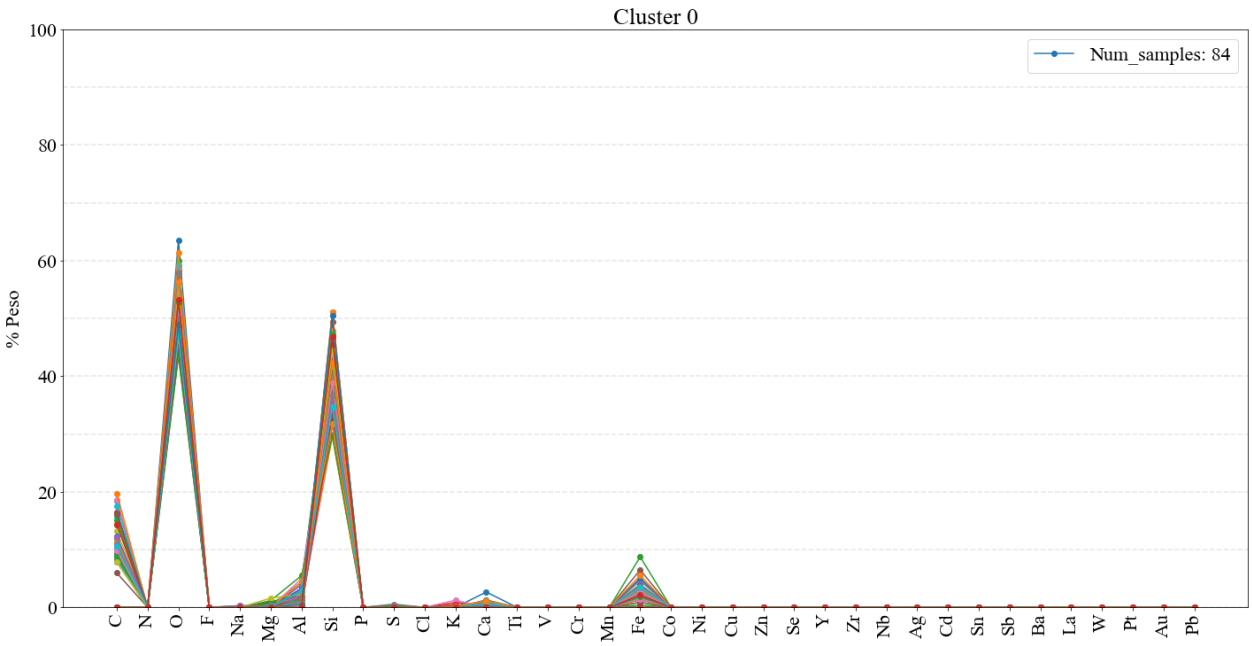
Anexo 36. Valle de Aburrá. Iteración 1. Clúster 14. Elaboración propia.



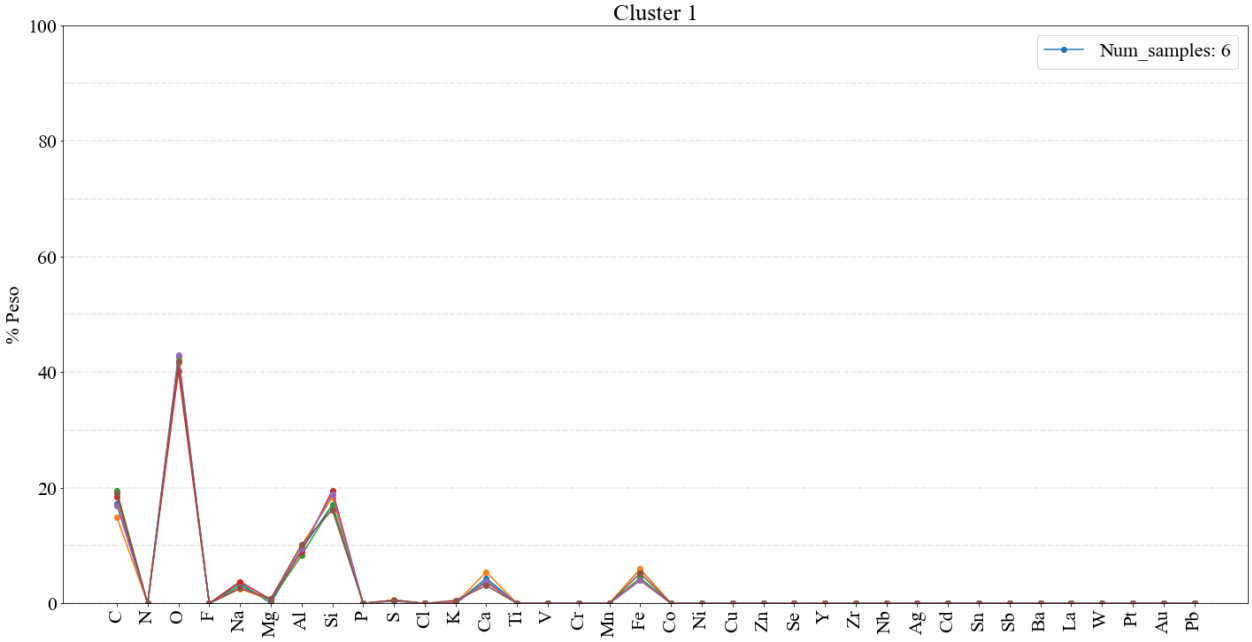
Anexo 37. Valle de Aburrá. Iteración 1. Clúster 15. Elaboración propia.



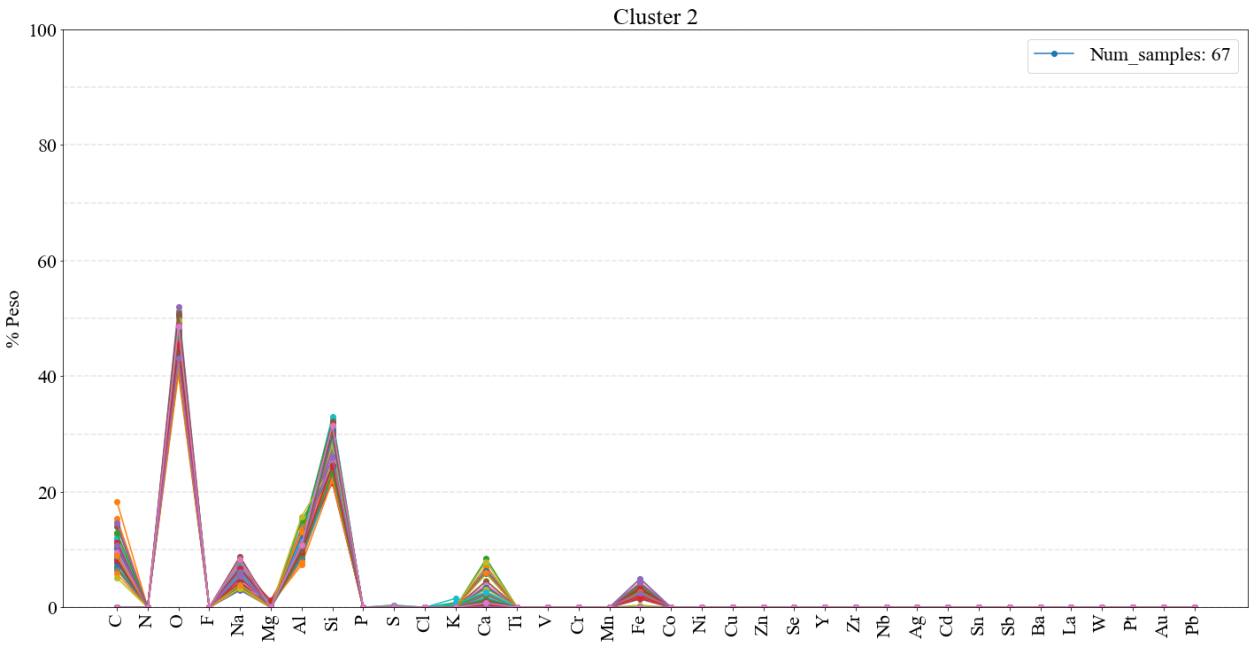
Anexo 38. Valle de Aburrá. Iteración 1. Clúster 16. Elaboración propia.



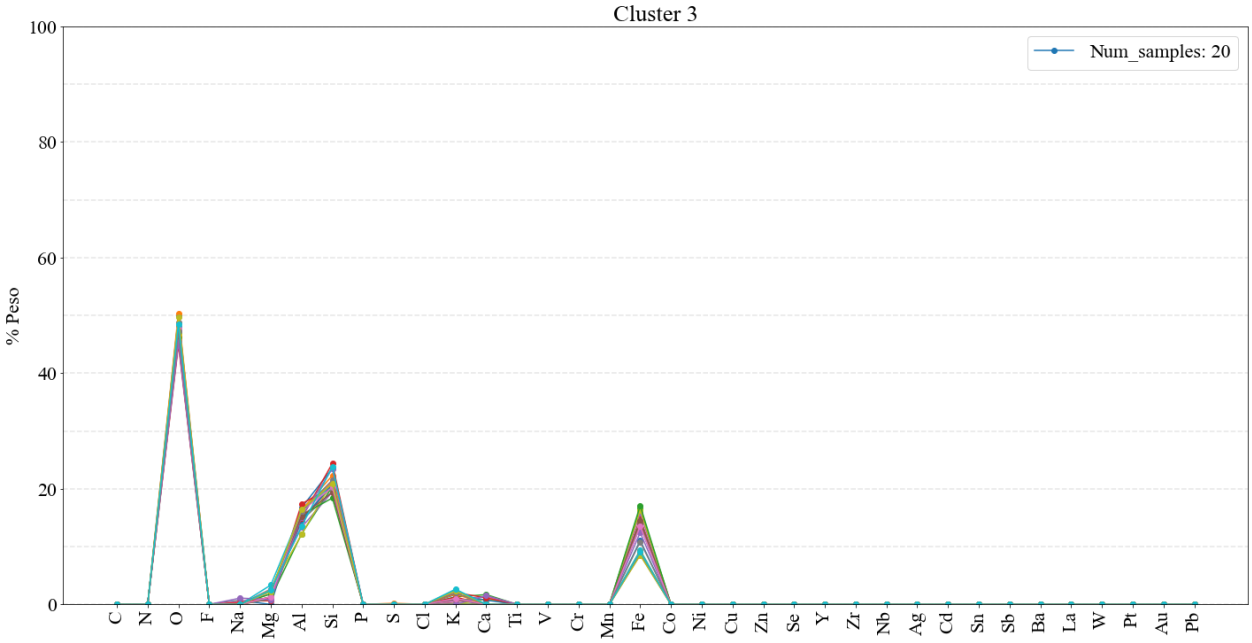
Anexo 39. Valle de Aburrá. Iteración 2. Clúster 0. Elaboración propia.



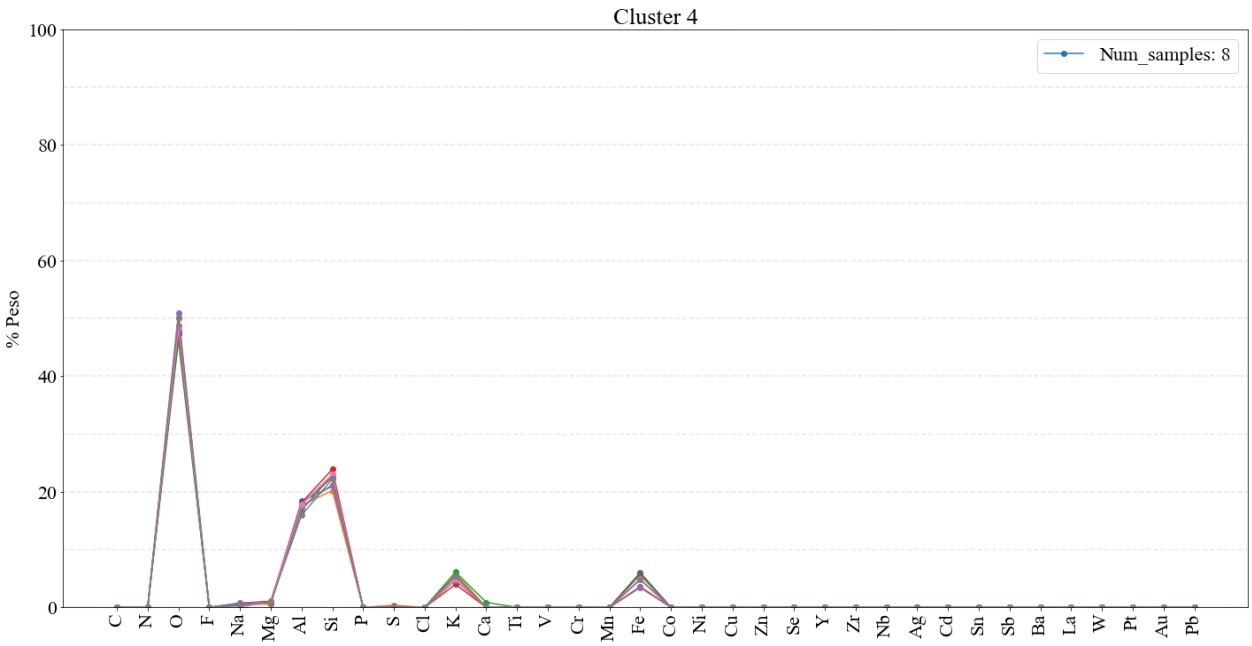
Anexo 40. Valle de Aburrá. Iteración 2. Clúster 1. Elaboración propia.



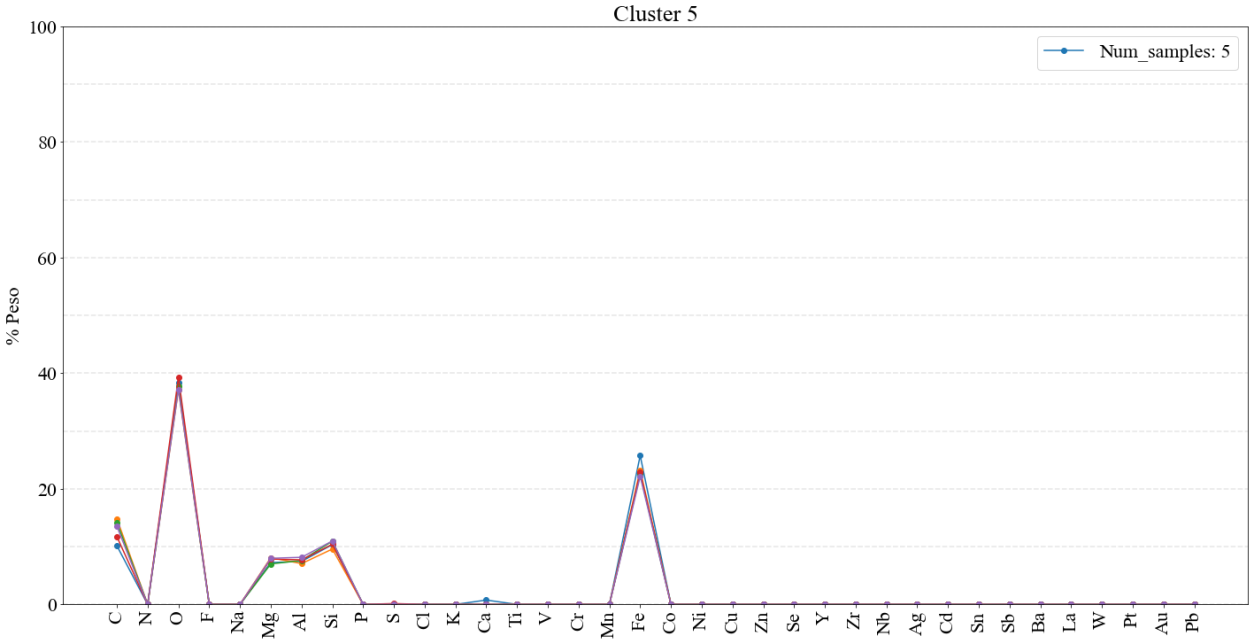
Anexo 41. Valle de Aburrá. Iteración 2. Clúster 2. Elaboración propia.



Anexo 42. Valle de Aburrá. Iteración 2. Clúster 3. Elaboración propia.

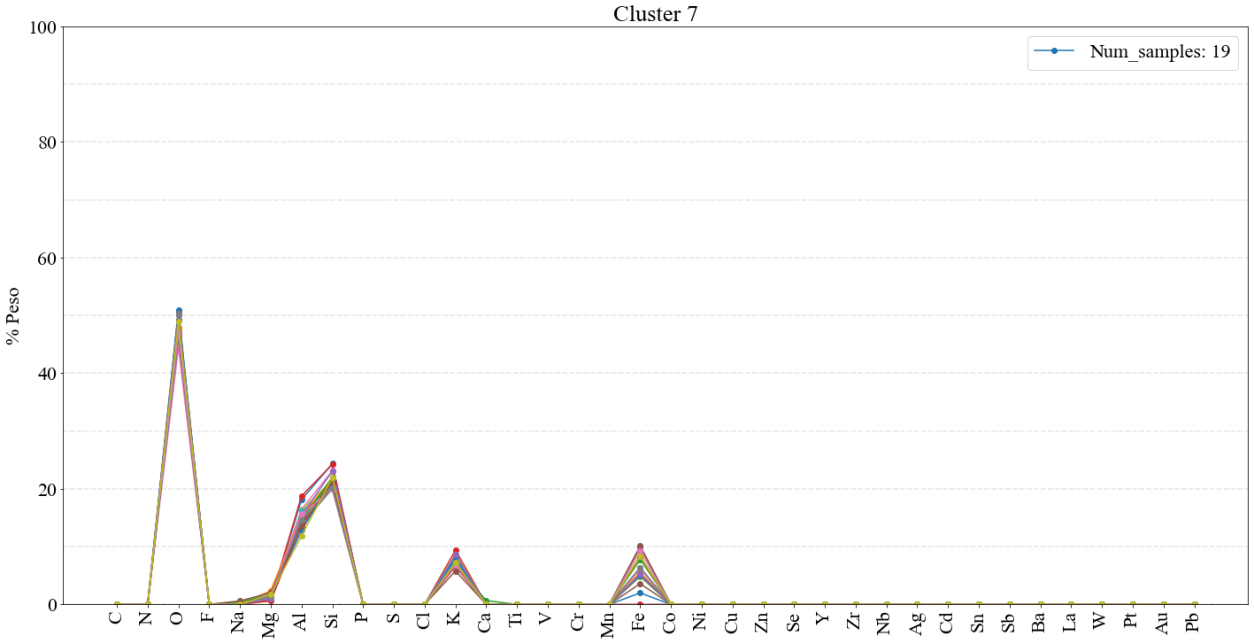


Anexo 43. Valle de Aburrá. Iteración 2. Clúster 4. Elaboración propia.

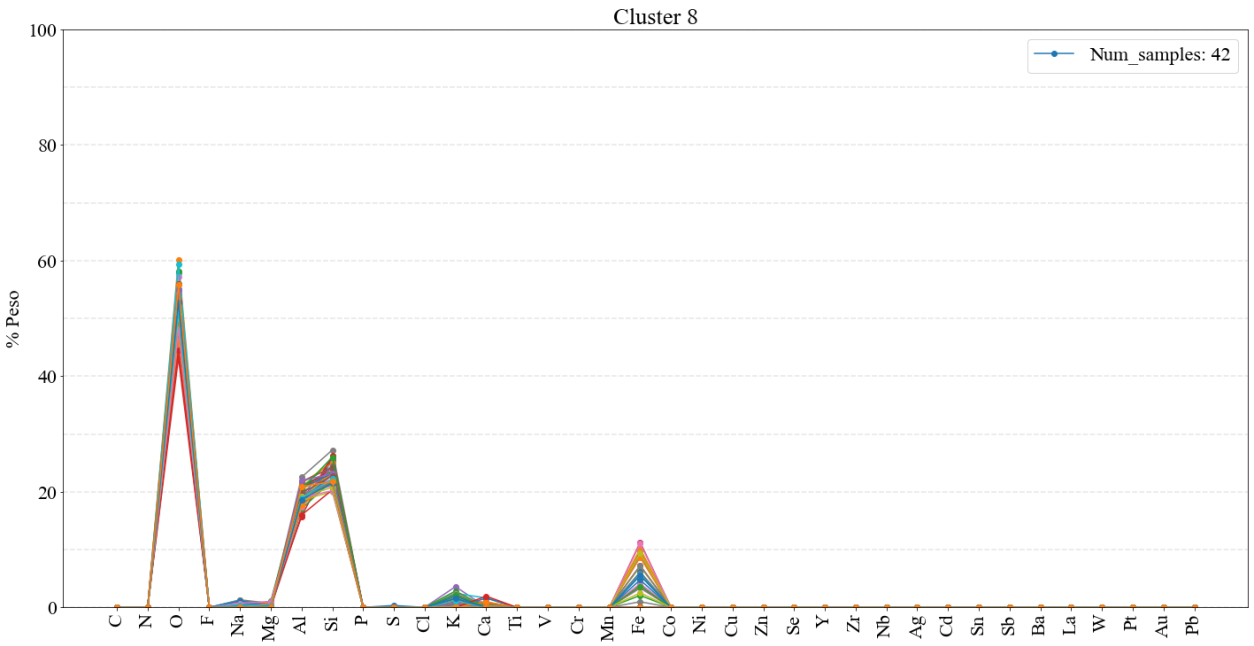


Anexo 44. Valle de Aburrá. Iteración 2. Clúster 5. Elaboración propia.

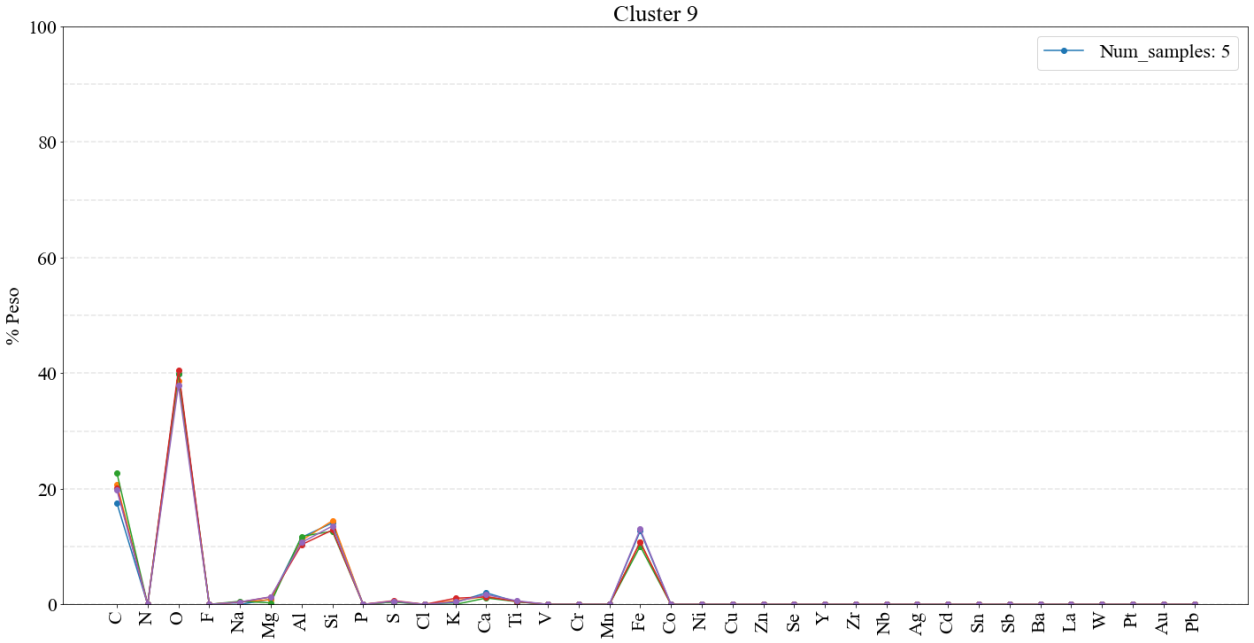
Anexo 45. Valle de Aburrá. Iteración 2. Clúster 6. Elaboración propia.



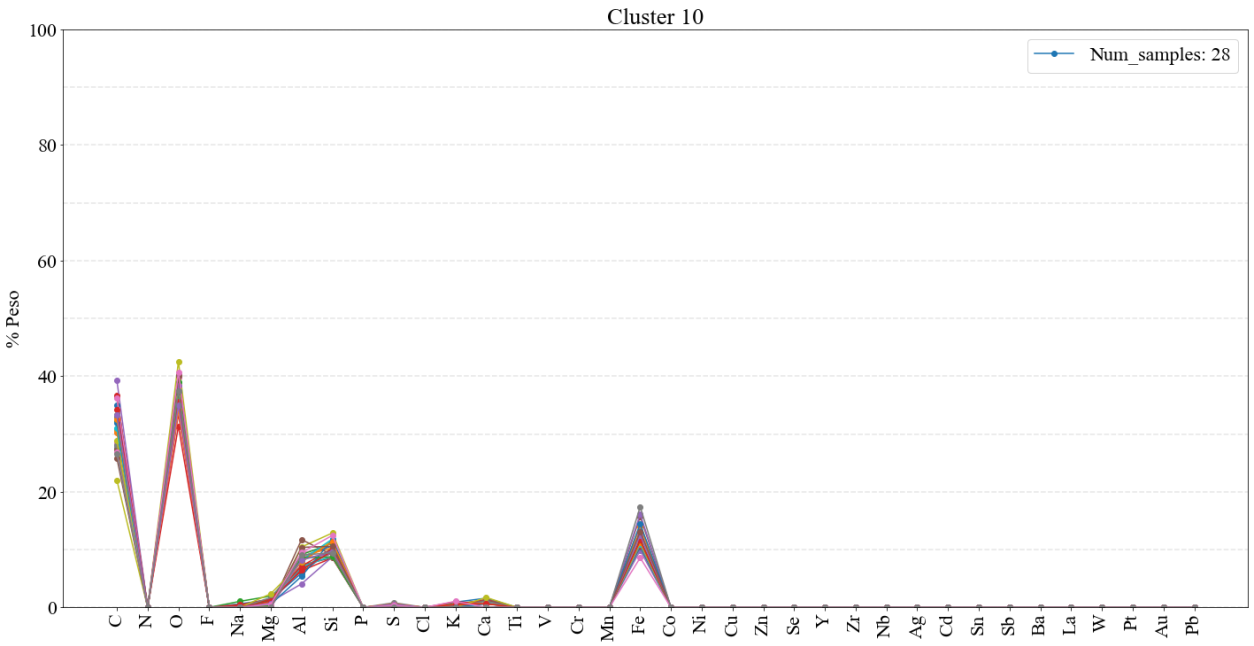
Anexo 46. Valle de Aburrá. Iteración 2. Clúster 7. Elaboración propia.



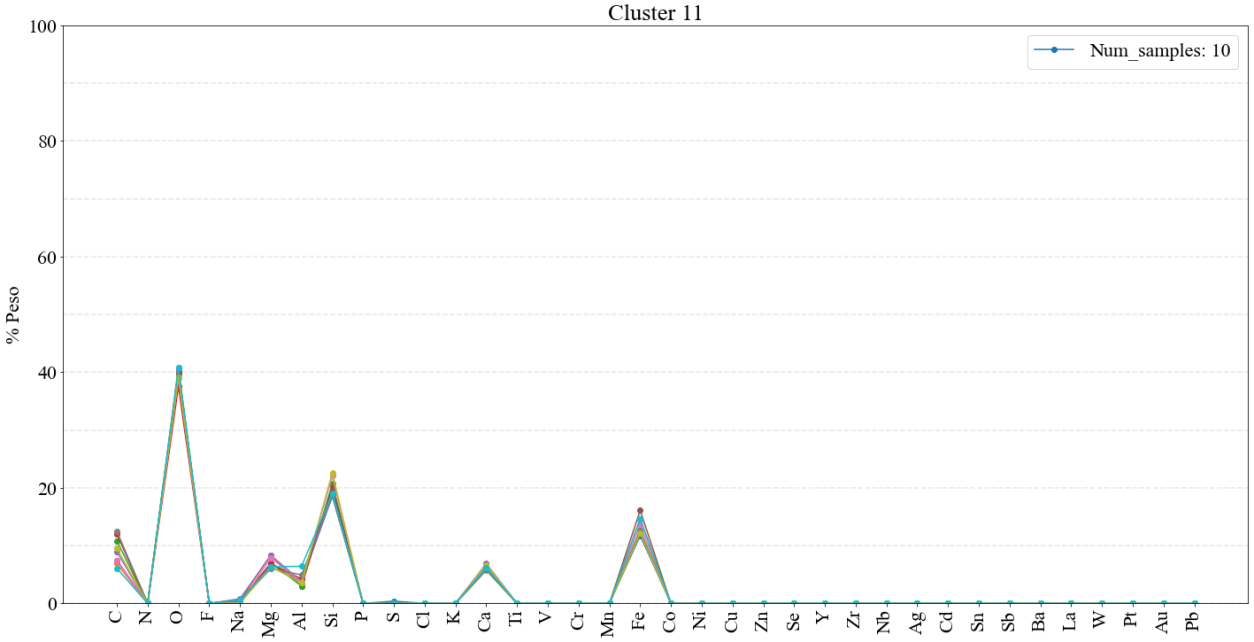
Anexo 47. Valle de Aburrá. Iteración 2. Clúster 8. Elaboración propia.



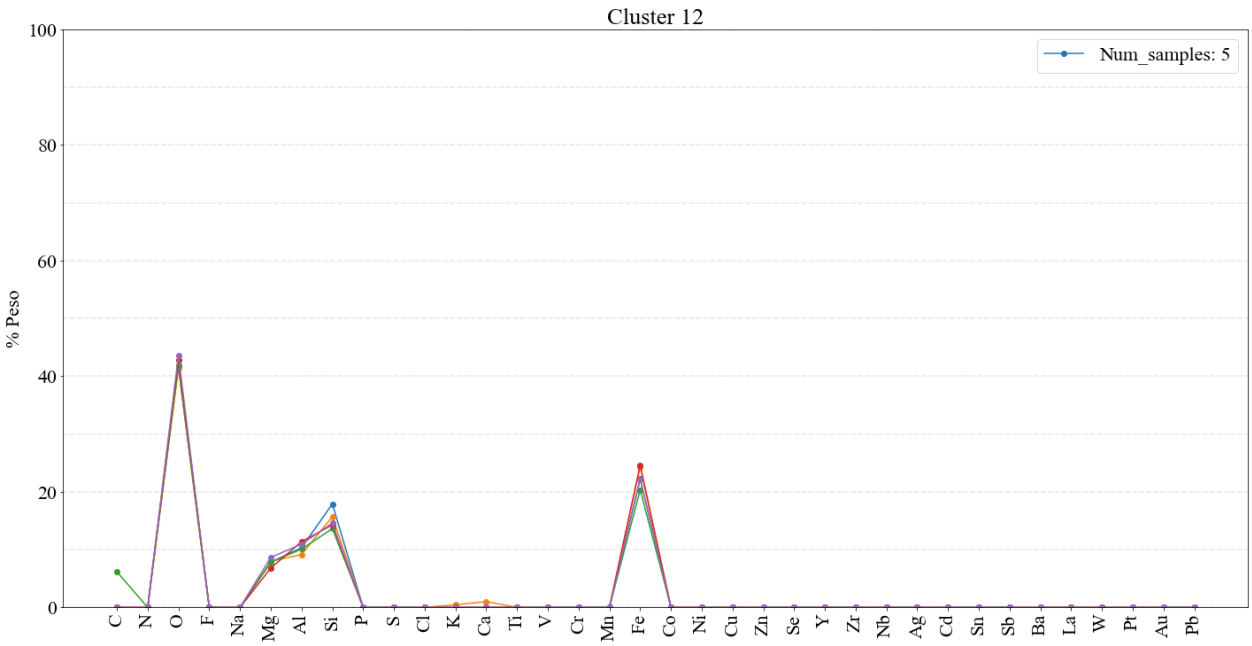
Anexo 48. Valle de Aburrá. Iteración 2. Clúster 9. Elaboración propia.



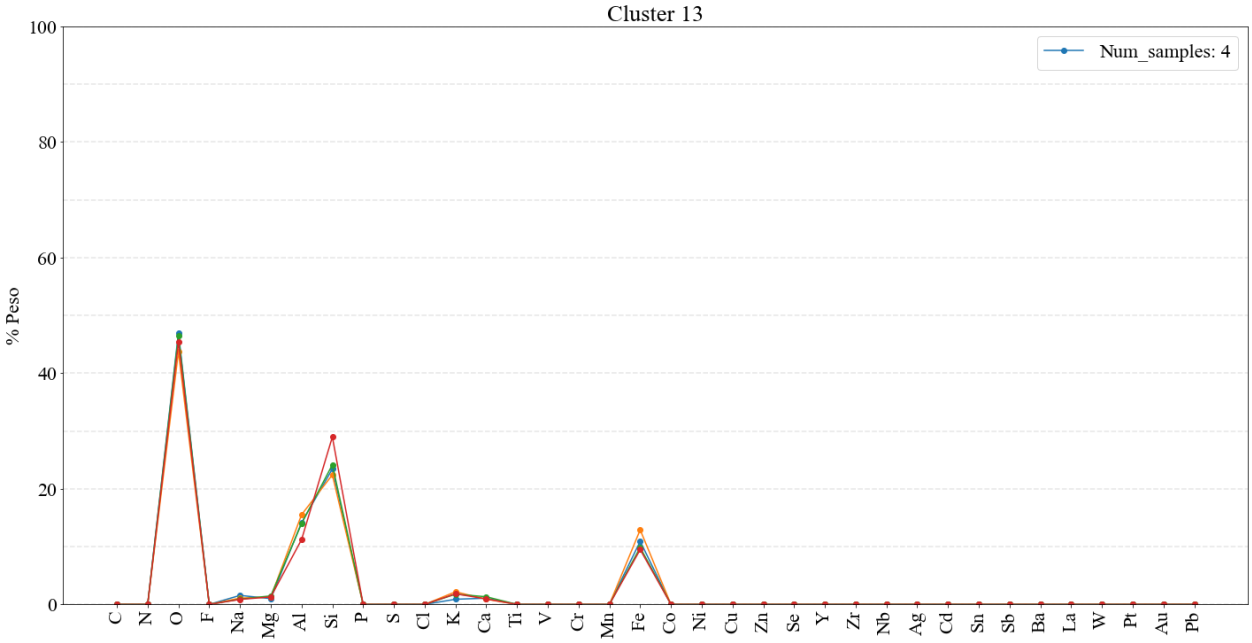
Anexo 49. Valle de Aburrá. Iteración 2. Clúster 10. Elaboración propia.



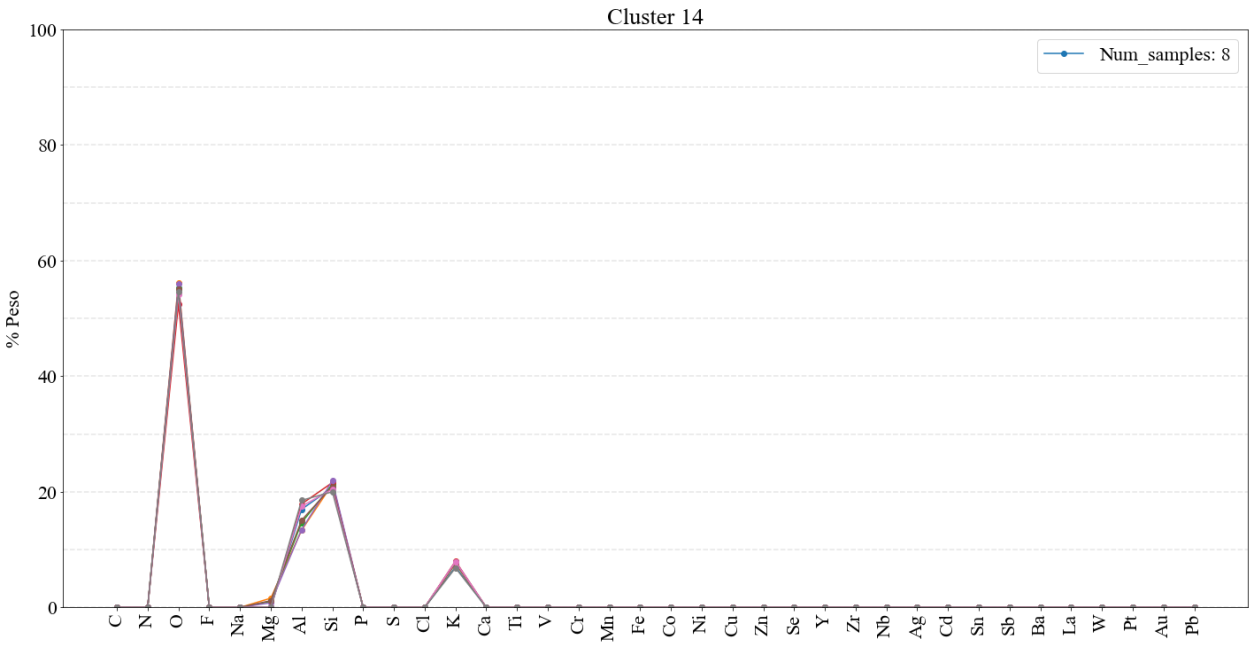
Anexo 50. Valle de Aburrá. Iteración 2. Clúster 11. Elaboración propia.



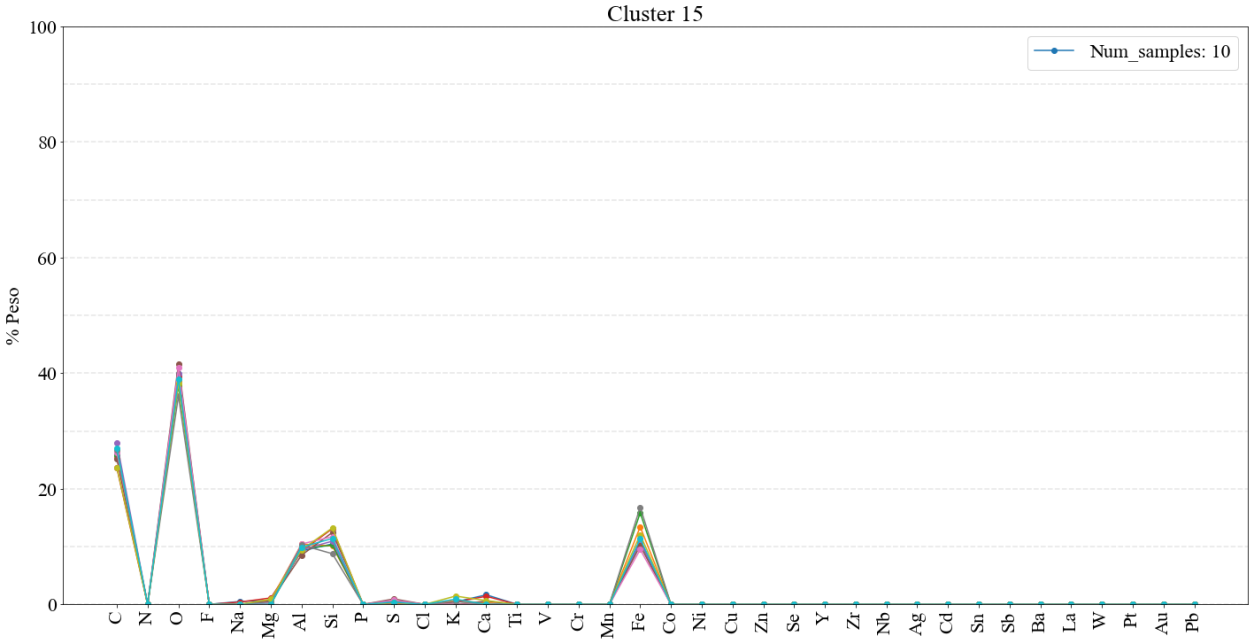
Anexo 51. Valle de Aburrá. Iteración 2. Clúster 12. Elaboración propia.



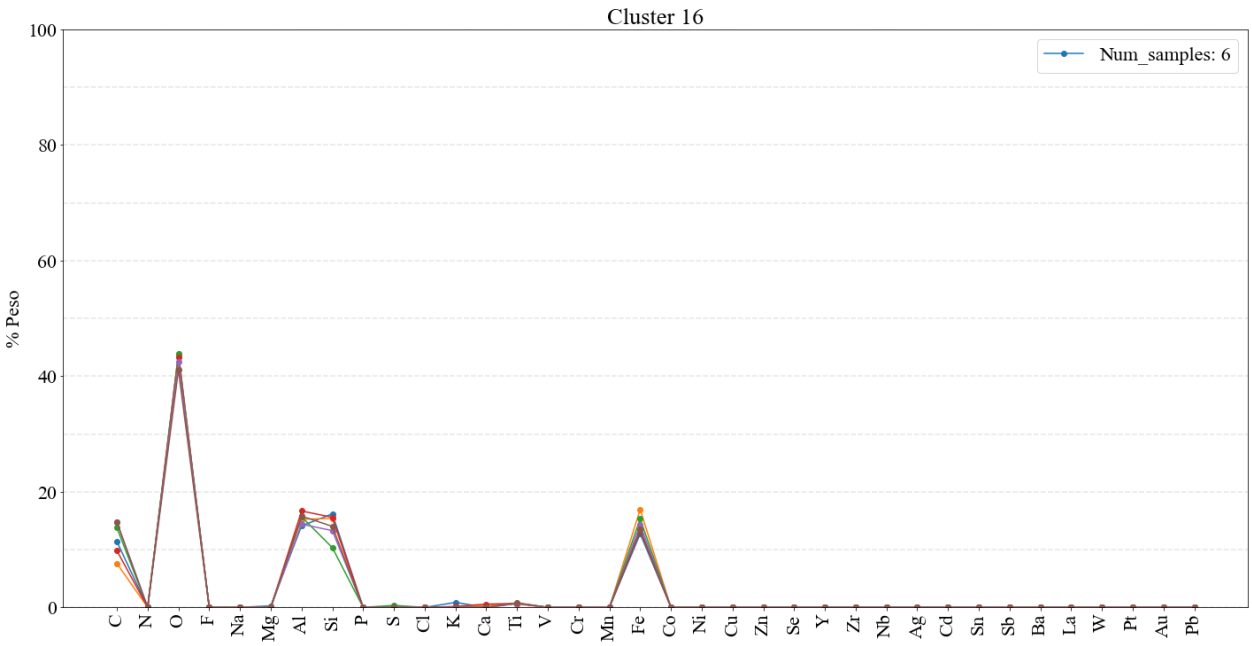
Anexo 52. Valle de Aburrá. Iteración 2. Clúster 13. Elaboración propia.



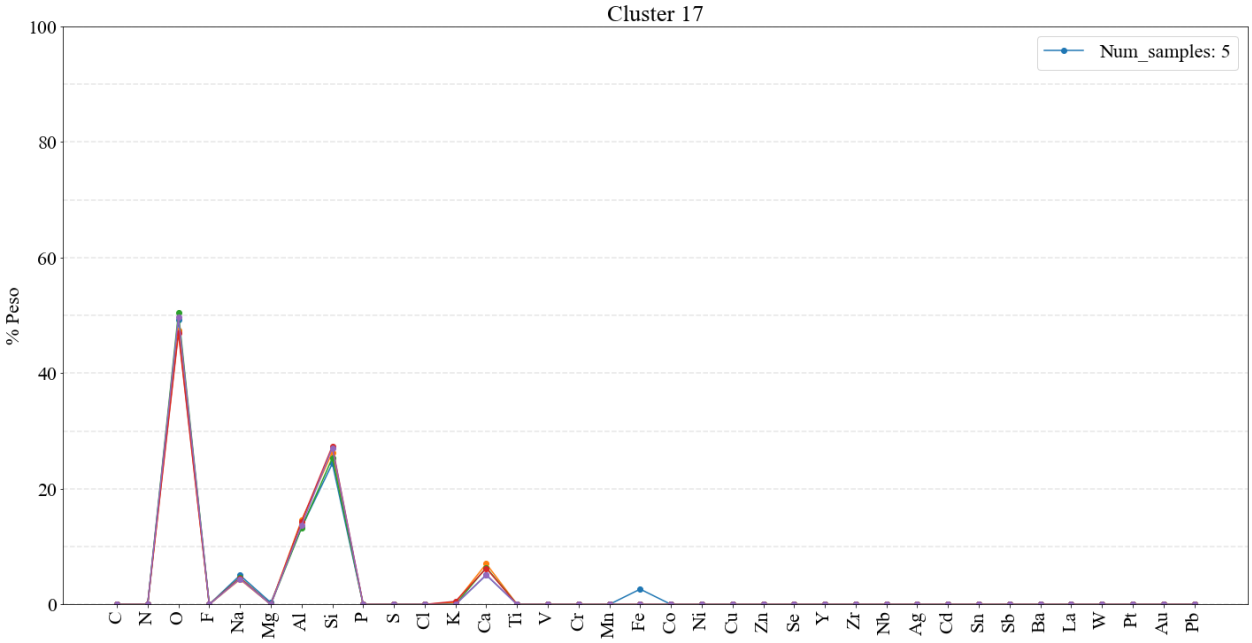
Anexo 53. Valle de Aburrá. Iteración 2. Clúster 14. Elaboración propia.



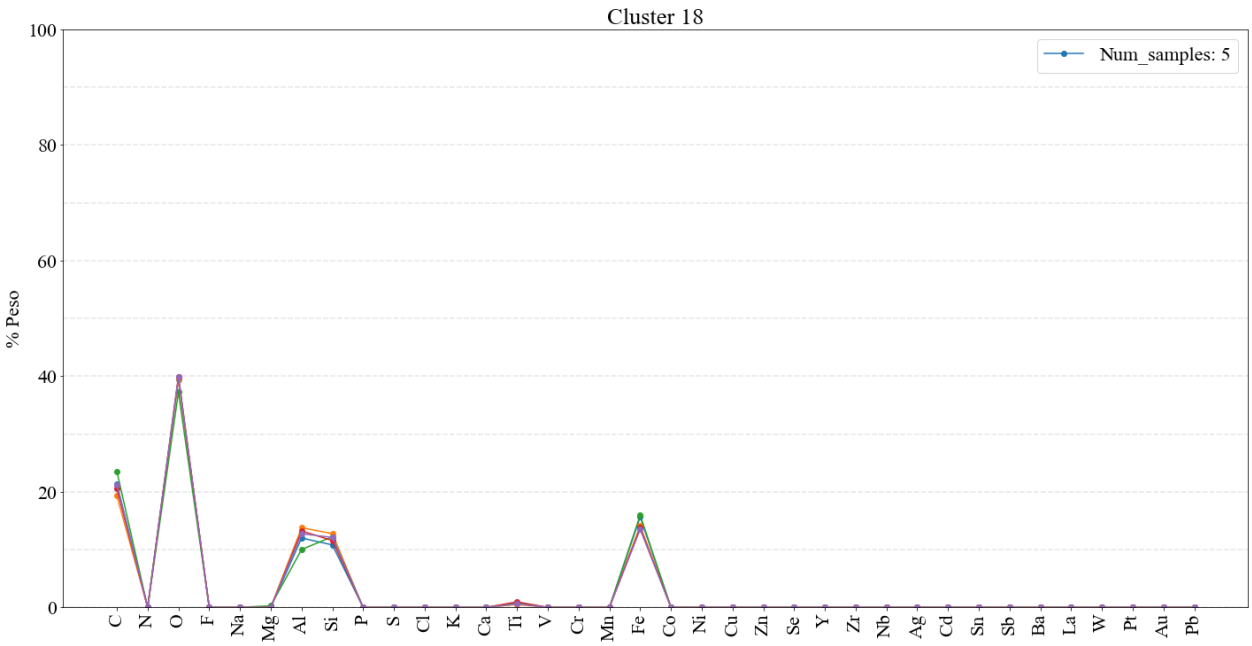
Anexo 54. Valle de Aburrá. Iteración 2. Clúster 15. Elaboración propia.



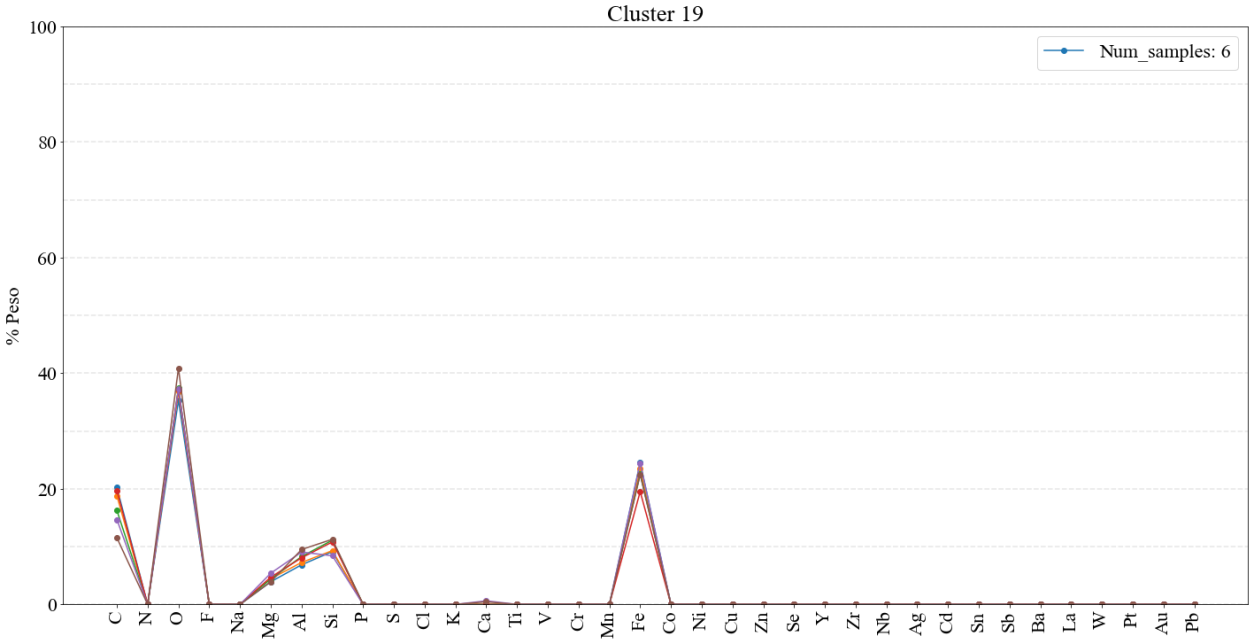
Anexo 55. Valle de Aburrá. Iteración 2. Clúster 16. Elaboración propia.



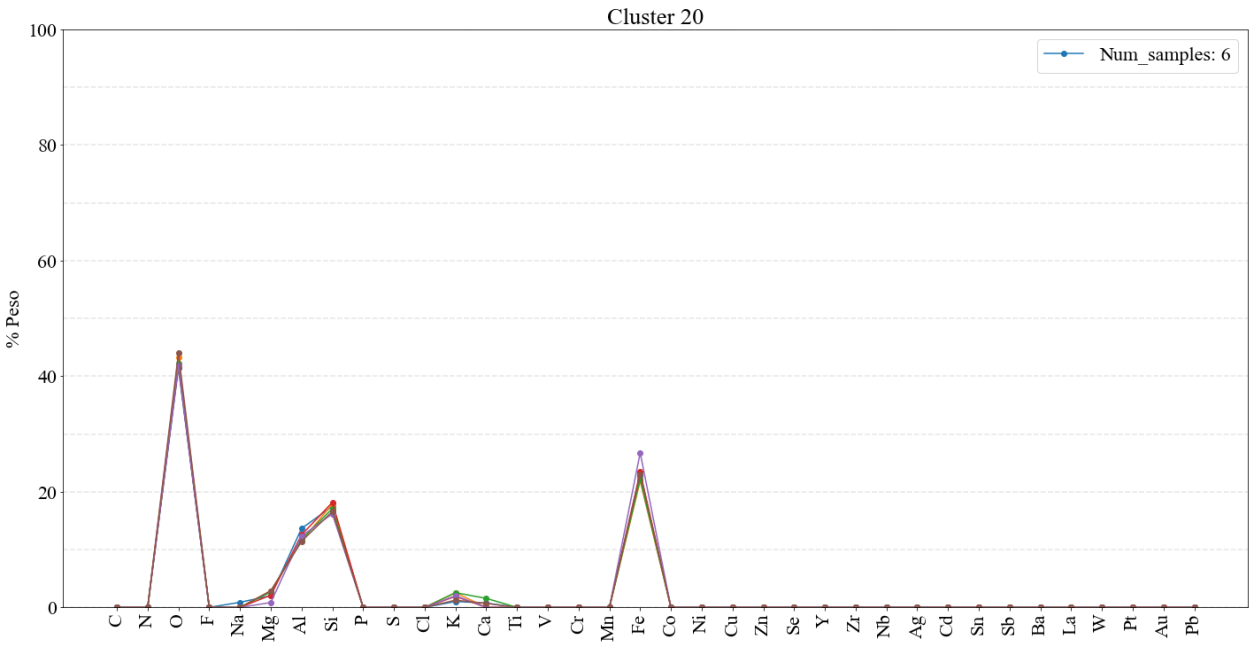
Anexo 56. Valle de Aburrá. Iteración 2. Clúster 17. Elaboración propia.



Anexo 57. Valle de Aburrá. Iteración 2. Clúster 18. Elaboración propia.



Anexo 58. Valle de Aburrá. Iteración 2. Clúster 19. Elaboración propia.



Anexo 59. Valle de Aburrá. Iteración 2. Clúster 20. Elaboración propia.

Tablas

Anexo 60. Definición de los grupos morfoquímicos. Tomada de Avellaneda et al. (2020).

Grupo morfo-químico	Definición	Fuente probable
Biogénicas-Orgánicas	Partículas principalmente compuestas por Carbono, Nitrógeno y/o Oxígeno	Esporas, polen, fragmentos de planta, materia orgánica
Mineral	Partículas que muestran una firma mineral (usualmente silicatos, calcita o mezclas)	Desgaste de la carreteras, trabajos de construcción, canteras, geogénico
Desgaste de llanta	Partículas mixtas compuestas con proporciones variables de caucho y minerales originados principalmente por la abración de la carretera. Estas partículas también pueden contener partículas metálicas (p.e. abración de los frenos)	Flujo vehicular
Óxidos metálicos	Partículas que están compuestas por metales (usualmente de Fe y Cu)	p.e. flujo vehicular y ferrocarril, industria

Anexo 61. Definición de parámetros morfológicos. Tomada de Avellaneda et al. (2020).

Parámetro	Descripción	Fórmula
Dimensión fractal	Parametro directamente relacionado con la rugosidad. Normalmete son calculadas dos dimensiones fractales D1 y D2. D1 y D2 describen la morfología (perímetro) y la textura (área) respectivamente. En este proyecto es unicamente utilizada D1. Cuando los valores de este parámetro son cercanos a uno, la partícula tiene una forma regular similar a un círculo	Existen múltiples algoritmos para calcular la dimensión fractal p.e. método de dilatación, dimensión Minkowski-Bouligand y dimensión caja
Convexidad	Parámetro susceptible a la rugosidad textural, cuando los valores son cercanos a uno, la partícula tiene un perímetro parecido al de un círculo	$\frac{Area_{particula}}{Area_{envolvente}}$
Solidez	Parámetro sensible a la rugosidad morfológica, cuando los valores son cercanos a uno, tiene un área parecida a la de un círculo	$\frac{Perimetro_{particula}}{Perimetro_{envolvente}}$
Circularidad	Parámetro susceptible a la forma y la rugosidad. Cuando el parámetro es más cercano a uno, su forma se asemeja a la de un círculo	$\frac{Perimetro_{particula}}{2\sqrt{\pi Area_{particula}}}$

Anexo 62. Base de datos de porcentaje en peso de elementos para minerales conocidos. Datos tomados de https://www.webmineral.com/. (consultada en 2023)

Mineral	H	C	O	F	Na	Mg	Al	Si	P	S	Cl	K	Ca	Ti	Cr	Mn	Fe	Cu	Zr	Ba	Nd	Hf	Pb
actinolita	0.24	0	45.01	0	0.59	9.71	1.39	25.64	0	0	0	0	8.6	0.11	0	0.13	8.58	0	0	0	0	0	0
albita	0	0	48.66	0	8.3	0	10.77	31.5	0	0	0	0	0.76	0	0	0	0	0	0	0	0	0	0
almandino	0	0	38.57	0	0	0	10.84	16.93	0	0	0	0	0	0	0	0	33.66	0	0	0	0	0	0
anortita	0	0	46.14	0	0.41	0	18.97	20.75	0	0	0	0	13.72	0	0	0	0	0	0	0	0	0	0
antigorita	1.34	0	47.88	0	0	18.18	0	18.68	0	0	0	0	0	0	0	0	13.93	0	0	0	0	0	0
apatito	0.07	0	38.76	1.24	0	0	0	0	18.25	0	2.32	0	39.36	0	0	0	0	0	0	0	0	0	0
augita	0	0	40.62	0	0.97	9.26	4.57	22.58	0	0	0	0	15.26	2.03	0	0	4.73	0	0	0	0	0	0
barita	0	0	27.42	0	0	0	0	0	0	13.74	0	0	0	0	0	0	0	0	0	58.84	0	0	0
biotita	0.41	0	43.36	1.1	0	14.02	6.22	19.44	0	0	0	9.02	0	0	0	0	6.44	0	0	0	0	0	0
calcita	0	12	47.96	0	0	0	0	0	0	0	0	0	40.04	0	0	0	0	0	0	0	0	0	0
calcopirita	0	0	0	0	0	0	0	0	0	34.94	0	0	0	0	0	0	30.43	34.63	0	0	0	0	0
caolinita	1.56	0	55.78	0	0	0	20.9	21.76	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cianita	0	0	49.37	0	0	0	33.3	17.33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
circon	0	0	33.63	0	0	0	0	14.76	0	0	0	0	0	0	0	0	0	0	43.14	0	3.78	4.69	0
corindón	0	0	47.07	0	0	0	52.93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
crisotilo	1.45	0	51.96	0	0	26.31	0	20.27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cromita	0	0	28.59	0	0	0	0	0	0	0	0	0	0	0	46.46	0	24.95	0	0	0	0	0	0
cuarzo	0	0	53.26	0	0	0	0	46.74	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
diopsido	0	0	44.33	0	0	11.22	0	25.94	0	0	0	0	18.51	0	0	0	0	0	0	0	0	0	0
dolomita	0	13.03	52.06	0	0	13.18	0	0	0	0	0	0	21.73	0	0	0	0	0	0	0	0	0	0
enstatita	0	0	47.81	0	0	24.21	0	27.98	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
epidota	0.19	0	40.05	0	0	0	3.9	16.22	0	0	0	0	15.44	0	0	0	24.2	0	0	0	0	0	0
ferrohornblenda	0.21	0	40.53	0	0	0	4.98	20.75	0	0	0	0	8.46	0	0	0	25.05	0	0	0	0	0	0
fluorita	0	0	0	48.67	0	0	0	0	0	0	0	0	51.33	0	0	0	0	0	0	0	0	0	0
galena	0	0	0	0	0	0	0	0	0	13.4	0	0	0	0	0	0	0	0	0	0	0	0	86.6
goethita	1.13	0	36.01	0	0	0	0	0	0	0	0	0	0	0	0	0	62.85	0	0	0	0	0	0
grafito	0	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
grosularia	0	0	42.62	0	0	0	11.98	18.71	0	0	0	0	26.69	0	0	0	0	0	0	0	0	0	0
hematita	0	0	30.06	0	0	0	0	0	0	0	0	0	0	0	0	0	69.94	0	0	0	0	0	0
illita	1.35	0	55.06	0	0	1.87	9.01	25.25	0	0	0	6.03	0	0	0	0	1.43	0	0	0	0	0	0
ilmenita	0	0	31.63	0	0	0	0	0	0	0	0	0	0	31.56	0	0	36.81	0	0	0	0	0	0
magnesiohornblenda	0.25	0	46.76	0	0	11.84	5.75	23.94	0	0	0	0	9.76	0	0	0	1.7	0	0	0	0	0	0
magnetita	0	0	27.64	0	0	0	0	0	0	0	0	0	0	0	0	0	72.36	0	0	0	0	0	0
malaquita	0.91	5.43	36.18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	57.48	0	0	0	0	0
montmorillonita	4.04	0	64.11	0	0.84	0	9.83	20.46	0	0	0	0	0.73	0	0	0	0	0	0	0	0	0	0
moscovita	0.46	0	47.35	0.95	0	0	20.3	21.13	0	0	0	9.81	0	0	0	0	0	0	0	0	0	0	0
ortoclasa	0	0	45.99	0	0	0	9.69	30.27	0	0	0	14.05	0	0	0	0	0	0	0	0	0	0	0
pirita	0	0	0	0	0	0	0	0	0	53.45	0	0	0	0	0	0	46.55	0	0	0	0	0	0
piropo	0	0	47.63	0	0	18.09	13.39	20.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
rutilo	0	0	40.06	0	0	0	0	0	0	0	0	0	0	59.94	0	0	0	0	0	0	0	0	0
tremolita	0.25	0	47.27	0	0	14.96	0	27.66	0	0	0	0	9.87	0	0	0	0	0	0	0	0	0	0
yeso	2.34	0	55.76	0	0	0	0	0	0	18.62	0	0	23.28	0	0	0	0	0	0	0	0	0	0

Anexo 63. Parámetros usados en las iteraciones para el agrupamiento

Ciudad	Iteración	Épsilon	Min pts.
Bogotá	1	1	5
Bogotá	2	0.8	5
Cali	1	1	5
Cali	2	1.8	5
Valle de Aburrá	1	0.8	5
Valle de Aburrá	2	0.85	5

Anexo 64. Código de DBSCAN para Jupyter Notebook (Python). Esta copia del código no conserva la división por celdas, se recomienda descargar el notebook para su uso. Elaboración propia.

Agrupación espacial basado en densidad de aplicaciones con ruido (DBSCAN)

DESCRIPCIÓN:

Gutiérrez-Silva, Juan Alberto (2023).

Basado en Ester et al. (1996) **(1)**

Algoritmo de agrupamiento o clustering espacial de múltiples usos.

Datos de entrada: Archivo CSV organizado de la siguiente manera:

Id muestra	Variable 1	Variable 2	Variable …	Variable n
Muestra 1	Valor 1	Valor 2	Valor …	Valor n
Muestra 2	Valor 1	Valor 2	Valor …	Valor n
Muestra …	Valor 1	Valor 2	Valor …	Valor n
Muestra n	Valor 1	Valor 2	Valor …	Valor n

Los valores deben estar separados por comas y usar puntos para marcar decimales.

(1) Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, No. 34, pp. 226-231).

COMENTARIOS:

Los gráficos que exporta este notebook están pensados para datos con más de 3 variables y variables que indiquen porcentajes (por eso la extensión de 0 a 100 en el eje Y, esta opción es personalizable). Para datos de 2 variables es más adecuado un scatterplot.

Para evitar fallos es recomendable limpiar el output antes de cada intento, reiniciar el kernel y correr el notebook completo.

Uso libre.

Importación de librerías

In [ ]:

```
#ES IMPORTANTE TENER PREVIAMENTE INSTALADAS LAS LIBRERÍAS

import os #interactuar con el sistema
import pandas as pd #análisis de datos
from sklearn.cluster import DBSCAN #clustering
from sklearn.preprocessing import StandardScaler
from sklearn.neighbors import NearestNeighbors
import matplotlib.pyplot as plt #gráficos
import tkinter as tk #ventanas emergentes
from tkinter import filedialog, simpledialog, messagebox
import numpy as np #cálculos
```

Importación de los datos

In [ ]:

```
root = tk.Tk()
root.withdraw()  # Ocultar la ventana principal de Tkinter

messagebox.showinfo("Información", "Seleccione el archivo CSV con los
datos.") # Crear una ventana emergente para seleccionar el archivo CSV

file_path = filedialog.askopenfilename(title="Selecciona el archivo CSV")
if not file_path:
messagebox.showwarning("Advertencia", "No se seleccionó ningún archivo.
Saliendo del programa.")
exit()

# Cargar los datos desde el archivo CSV

data = pd.read_csv(file_path, engine='python', index_col=0) #con
index_col=0 se indica que las columnas tienen nombre
```

Ver tipo de datos

In [ ]:

```
data.info()
data.head()
```

Variables a analizar

In [ ]:

```
# Solicitar el número de variables a analizar

num_variables = simpledialog.askinteger("Número de Variables", "Introduce
el número de variables:")

if num_variables is None:
messagebox.showwarning("Advertencia", "No se proporcionó el número de
variables. Saliendo del programa.")
exit()

X = data.iloc[:, :num_variables]
```

Parametrización de DBSCAN

In [ ]:

```
# Normalizar los datos

scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

In [ ]:

```
# Solicitar al usuario el número de puntos vecinos

Neighbors = simpledialog.askinteger("Neighbors", "Introduce el número de
puntos vecinos para la estimación de epsilon:")
neigh = NearestNeighbors(n_neighbors=Neighbors)
nbrs = neigh.fit(X_scaled)
distancias, indices = nbrs.kneighbors(X_scaled)

distancias = np.sort (distancias, axis=0) #ordena las distancias de menor a
mayor
dist= Neighbors-1
distancias = distancias [:,dist] #selecciona la última distancia, es decir
la mayor

#Crear gráfico de codo

fig = plt.figure(figsize=(7,7))
plt.plot(distancias)
plt.xlabel("puntos")
plt.ylabel("distancia")

# Agregar líneas guía al gráfico

espaciado = 1
min_dist = np.min(distancias)
max_dist = np.max(distancias)
for i in range(int(min_dist), int(max_dist), espaciado):
plt.axhline(i, color='lightgray', linestyle='--', linewidth=0.5)

plt.show()
print(f'distancia: {dist}')
```

In [ ]:

```
# Solicitar al usuario el valor epsilon

Epsilon = simpledialog.askfloat("Epsilon", "Introduce el valor de epsilon ,
puedes usar el grafico de codo como apoyo:")

if num_variables is None:
print("No se proporcionó el número de variables.")
exit()

min_samples = Neighbors
```

Nota :Epsilon (ε) determina la distancia máxima entre dos puntos de datos para que se consideren vecinos cercanos, ε define un radio alrededor de cada punto de datos central, y todos los puntos dentro de ese radio se consideran vecinos del punto central.

Clustering DBSCAN

In [ ]:

```
# Crear un objeto DBSCAN y ajustarlo a los datos

dbscan = DBSCAN(eps=Epsilon, min_samples=min_samples)  #Usa los datos de ε
y minimo de vecinos solicitados antes
dbscan.fit(X_scaled)
```

In [ ]:

```
# Obtener las etiquetas de clúster asignadas a cada punto de datos

cluster_labels = dbscan.labels_

# Agregar la columna de clústeres al DataFrame original

data['Cluster'] = cluster_labels
```

In [ ]:

```
# Contar el número de clusters y puntos de ruido (-1 representa ruido)

n_clusters = len(set(cluster_labels)) - (1 if -1 in cluster_labels else 0)
n_noise = list(cluster_labels).count(-1)

# Mostrar el número de clusters y puntos de ruido

messagebox.showinfo("Información",f"Número de clusters encontrados:
'{n_clusters}'.")
messagebox.showinfo("Información",f"Número de puntos de ruido:
'{n_noise}'.")

# Imprimir el número de clusters y puntos de ruido

print(f'Número de clusters encontrados: {n_clusters}')
print(f'Número de puntos de ruido: {n_noise}')
```

Creación de los gráficos

In [ ]:

```
# Transponer los datos para crear un gráfico de líneas

cluster_data = X[cluster_labels != -
1].groupby(cluster_labels[cluster_labels != -1])
fig, ax = plt.subplots(figsize=(14, 5))

messagebox.showinfo("Información", "Seleccione una carpeta para guardar los
gráficos.")

# Solicitar al usuario la carpeta para guardar los gráficos

output_folder = filedialog.askdirectory(title="Selecciona la carpeta para
guardar los gráficos")
```

```
if not output_folder:
messagebox.showwarning("Advertencia", "No se seleccionó ninguna carpeta.")
exit()

#tipografía del gráfico

from matplotlib import rcParams
rcParams['font.family'] = 'serif'  # Puedes cambiar 'serif' por el nombre
de la fuente que desees
rcParams['font.serif'] = ['Times New Roman']  # Puedes especificar una
fuente específica aquí
rcParams['font.size'] = 22  # Establece el tamaño de fuente predeterminado

# Crear gráficos de líneas para cada clúster y guardar imágenes

unique_clusters = set(cluster_labels)
for cluster in unique_clusters:
if cluster == -1:
      continue  # Saltar el cluster de ruido (-1)

cluster_data = X[cluster_labels == cluster]
cluster_data_transposed = cluster_data.T  # Transponer los datos para que
las variables estén en el eje x

# Configuración del gráfico
fig, ax = plt.subplots(figsize=(24, 12))
ax.plot(cluster_data_transposed, marker='o')

# Establecer límites del eje vertical de 0 a 100
ax.set_ylim(0, 100)

# Etiquetas de eje y
ax.set_ylabel('% Peso')

# Etiquetas de eje x (nombres de variables)
ax.set_xticks(range(num_variables))
ax.set_xticklabels(X.columns, rotation=90)

# Título del gráfico
ax.set_title(f'Cluster {cluster}')

# Líneas guía verticales
for y in range(0, 101, 10):
      ax.axhline(y, color='gray', linestyle='--', alpha=0.2)

# Calcular la cantidad de datos en el cluster actual
num_data_points = len(cluster_data)

# Agregar leyenda con la cantidad de datos
ax.legend([f'Num_samples: {num_data_points}'], loc='upper right')

# Guardar la imagen en la carpeta especificada
output_filename = os.path.join(output_folder, f'cluster_{cluster}.png')
```

```
plt.savefig(output_filename, bbox_inches='tight')
plt.show()
```

Agregando la clasificación y exportando el archivo CSV con clústeres.

In [ ]:

```
# Solicitar al usuario la ubicación y el nombre del archivo para guardar
los resultados

messagebox.showinfo("Información", "Seleccione una carpeta y un nombre para
guardar los datos agrupados.")

output_file = filedialog.asksaveasfilename(title="Guardar archivo CSV",
defaultextension=".csv")
if not output_file:
messagebox.showwarning("Advertencia", "No se seleccionó ninguna carpeta.")
exit()
if output_file:
data.to_csv(output_file, index= True)
messagebox.showinfo("Información", f"Los resultados se han guardado en
'{output_file}'.")
```

In [ ]:

```
#Mensaje de aviso al final (ventana emergente)
messagebox.showinfo("Información", "¡EL PROCESO HA FINALIZADO!")
```