Generación de un modelo numérico de agrupamiento (Clustering) para la identificación de microplásticos recolectados en el Parque Nacional Natural Corales de Profundidad y en el Parque Nacional Natural Los Corales del Rosario y San Bernardo, mediante Espectroscopía Infrarroja por Transformada de Fourier con Reflectancia Total Atenuada (FTIR-ATR).

Diego F. Jaimes Castro [1] e Isabel C. Prada Buitrago [2]

Trabajo de grado como requisito para optar a los títulos de Físico[1] e Ingeniera Química[2]

Director:

Jader Enrique Guerrero Bermúdez

Doctor en Ciencias Naturales (Física)

Codirector:

Rafael Cabanzo Hernández

Magíster en Física

Universidad Industrial de Santander

Facultad de Ciencias

Escuela de Física

Bucaramanga

2023

Dedicatoria

Dedicado a mi adorada Ana Sofía, mi pequeña fuente de luz y felicidad, me has enseñado a redescubrir el mundo a través de tus ojos, convirtiendo lo ordinario en algo mágico y extraordinario. Con amor y determinación podemos alcanzar nuestros objetivos más grandes. Nunca temas perseguir tus sueños, tal como yo lo hice mientras te sostenía en mis brazos. Lo logramos, hija.

A mi esposa Carolina, has sido mi motivación y fortaleza en los días más difíciles de este largo viaje. Sin tu amor y tu apoyo incondicional, esto no habría sido posible. Gracias por creer en mi en cada momento. Lo logramos, amor.

A mi madre Fidelina, por todo el esfuerzo que hiciste para asegurarte que tuviera todo lo que necesitaba, aún tras la pérdida de papá y las condiciones adversas. Eternamente gracias. Lo logramos, papás.

Diego F. Jaimes Castro

Dedicatoria

A mi familia, por su amor, paciencia y comprensión, y la bondad de sus corazones al entender sin cuestionar.

A Blanca Díaz, Elkin Castro, Karla Corredor e Ingrid Mesa, mi agradecimiento más grande y profundo por mantenerme en pie y caminar a mi lado cuando el miedo, la duda y la incertidumbre no dejaban cabida para la esperanza.

A la de 22, con todo mi amor, cumpliéndole lo que le debía.

Isabel C. Prada Buitrago

Agradecimientos

Mi más sincero agradecimiento al doctor Jader Guerrero, nuestro director de proyecto, su orientación experta, paciencia infinita y dedicación incansable hicieron posible finalizar esta tesis. De la misma manera, agradezco al codirector del proyecto, doctor Rafael Cabanzo, por sus valiosos aportes en el desarrollo del mismo, me siento afortunado de haber tenido la oportunidad de trabajar bajo la guía y liderazgo de ustedes.

A la Vicerrectoría de investigación y extensión (VIE) de la Universidad Industrial de Santander agradecemos el financiamiento del macroproyecto 2839 del que hace parte la presente investigación, proyecto 'Evaluación de la contaminación por microplásticos sobre la comunidad de zooplancton en el Parque Nacional Natural Corales de Profundidad y en el Parque Nacional Natural Corales del Rosario y San Bernardo'.

Al Laboratorio de Espectroscopía Atómica y Molecular, y en especial a Ximena Calderón, por su apoyo constante. Asi mismo, al Laboratorio de Hidrobiología y a Parques Nacionales por su invaluable labor, y a mi compañera Isabel Prada. Sin la ayuda de todos ustedes, no habría sido posible la culminación de este proyecto.

Diego F. Jaimes Castro

Agradecimientos

Agradecimientos a la Vicerrectoría de Investigación y Extensión de la Universidad Industrial de Santander, por el financiamiento del macroproyecto del que hizo parte nuestra investigación; al Laboratorio de Espectroscopía Atómica y Molecular, al Laboratorio de Hidrobiología y a Parques Nacionales por permitirnos contar con la invaluable colaboración, conocimientos y experticia de su talento humano.

Quisiera tomar este último párrafo a modo más personal y pido me disculpen quienes lo lean y lo encuentren demasiado informal, créanme, es con el mayor respeto y sinceridad que escribo estas palabras. Mis más sinceros agradecimientos al doctor Jader Guerrero, el director de proyectos más amable, generoso y enérgico; a los doctores Rafael Cabanzo, el más cordial y afable codirector de proyectos, y Enrique Mejía, inmensamente agradecida por darme la gran oportunidad de participar en este proyecto; ellos no lo saben, pero cuando fui por primera vez a conocerlos, los nervios de que dijeran que no apenas me dejaban hablar.

A Ximena Calderón, por siempre tener la mejor disposición para orientarme y ayudarme en todo lo que fuera necesario, y por garantizar el casi inagotable suministro de café, que nunca falte su aroma en el laboratorio. A Alejandra Ruiz, por responder a todas mis preguntas, importantes y simples, con tanta amabilidad y el mejor sentido del humor. Y finalmente a mi compañero de trabajo de grado, Diego Jaimes, mis más sinceras gracias, Diego, gracias por todo.

Isabel C. Prada Buitrago

Tabla de contenido

Resumen **11**

1 Marco teórico **16**

1.1 Los microplásticos en los océanos . . . 16

1.2 Espectroscopía infrarroja por transformada de Fourier - reflexión total atenuada. (FTIR - ATR) 17

1.3 Fundamentos del tratamiento de datos: Reducción de dimensionalidad mediante PCA (Análisis de componentes principales). . . . 18

1.4 Modelos de agrupamiento. . . . 20

1.4.1 Modelos de aprendizaje no supervisado. . . . 20

1.4.2 *k-means.* . . . 21

1.4.3 Agrupamiento jerárquico. . . . 21

1.5 Índices de validación interna . . . 22

1.5.1 Índice de Dunn . . . 23

1.5.2 Índice de Davies-Bouldin . . . 23

1.5.3 Coeficiente de siluetas simplificado . . . 24

2 Metodología **25**

2.1 Obtención de muestras . . . 25

2.2 Entrega de muestras . . . 27

2.3 Protocolo de muestreo . . . 27

2.3.1 Parámetros instrumentales . . . 28

3 Análisis y discusión de resultados **29**

3.1 Dimensionalidad obtenida mediante PCA . . . 29

3.2 Pretratamiento del conjunto de datos para *k-means* . . . 32

3.3 Resultados del agrupamiento *k-means* . . . 36

3.4 Resultados del agrupamiento por enfoque jerárquico: dendrograma . . . 38

3.5 Confrontación con base de datos . . . 41

3.5.1 Grupos sugeridos por *k-means* . . . 41

3.5.2 Grupos sugeridos por enfoque jerárquico aglomerativo 46
3.6 Distribución de las muestras en las estaciones de los PNNs 50

4 Conclusiones **53**

5 Consideraciones y estudios posteriores **54**

Bibliografía **55**

A Anexos **61**
A.1 Anexo A. Índice de Dunn 61
A.2 Anexo B. Índice de Davies - Bouldin 62
A.3 Anexo C. Coeficiente Silueta simplificado 63
A.4 Anexo D. Códigos implementados en Matlab R2019b 64

Lista de Figuras

Figura 1: Esquema de montaje óptico de un espectrómetro *ATR - FTIR*. 18

Figura 2: Ubicación geográfica de las estaciones de medición en los parques nacionales naturales Los Corales del Rosario y San Bernardo, y Corales de Profundidad. 26

Figura 3: Las muestras más comunes fueron fibras laminares y cilíndricas como las mostradas a la izquierda y derecha de la figura, respectivamente. 27

Figura 4: Representación de la varianza acumulada en función del número de componentes principales. La señal se explica por encima del 90 % al considerar más de siete componentes principales. 29

Figura 5: Las siete primeras componentes principales conforman una base vectorial, de esta manera cada espectro se expresa mediante los siete coeficientes que ponderan estos vectores propios. 30

Figura 6: Comportamiento del índice de Dunn para diferentes propuestas de modelo de agrupamiento, en función del número de grupos y número de componentes principales. En **(a)**, **(e)** e **(i)**, se consideran siete (7), once (11) y quince (15) componentes principales. 33

Figura 7: Característica del índice de Davies-Bouldin para diferentes propuestas de modelo de agrupamiento. En **(a)**, **(e)** e **(i)**, se consideran siete (7), once (11) y quince (15) componentes principales. Nótese su tendencia contraria al coeficiente de Dunn, a medida que aumenta el número de grupos. 34

Figura 8: Curva de coeficiente silueta versus número de grupos. En **(a)**, **(e)** e **(i)**, se consideran siete (7), once (11) y quince (15) componentes principales. 35

Figura 9: Valor silueta para un modelo de cinco grupos, atendiendo al mejor índice de Dunn. . 37

Figura 10: La propuesta de cinco grupos sugiere los espectros promedio mostrados desde (a) hasta (e). Estos espectros promedio son instrumentos para explorar la presencia de MPs en las estaciones de los PNN considerados. 37

Figura 11: El dendrograma es la síntesis del enfoque aglomerativo. La línea de trazo permite discriminar cinco grupos (cantidad que sugieren los índices de validación interna). Debe observarse la similaridad existente entre la envergadura de las ramas en el dendrograma y el perfil de siluetas (Figura 9). 39

Figura 12: Para efecto de comparar con la estrategia *k-means* se corta el dendrograma de manera que resulten cinco grupos. En negrilla se presenta el efecto promedio. 40
Figura 13: Adaptación de los espectros promedio calculados por la estrategia *k-means* con espectros sugeridos por la base de datos *KnowItAll*. 44
Figura 14: Adaptación de los espectros promedio sugeridos por el enfoque aglomerativo jerárquico de la base de datos *KnowItAll*. 48

Figura 15: Curva de índices de Dunn versus número de grupos. En (b), (c), (d), (f), (g) y (h), se consideran ocho (8), nueve (9), diez (10), doce (12), trece (13) y catorce (14) componentes principales, respectivamente. 61
Figura 16: Curva de índices de Davies - Bouldin versus número de grupos. En (b), (c), (d), (f), (g) y (h), se consideran ocho (8), nueve (9), diez (10), doce (12), trece (13) y catorce (14) componentes principales, respectivamente. 62
Figura 17: Curva de coeficiente de siluetas versus número de grupos. En (b), (c), (d), (f), (g) y (h), se consideran ocho (8), nueve (9), diez (10), doce (12), trece (13) y catorce (14) componentes principales, respectivamente. 63

Lista de Tablas

Tabla 1: Estaciones de muestreo para la exploración de microplásticos en los Parques Nacionales Naturales (PNN) Corales del Rosario y San Bernardo (CRSB) y Corales de Profundidad (CPR). 26

Tabla 2: Los espectros de las muestras colectadas fueron registrados con los siguientes parámetros instrumentales. 28

Tabla 3: Índices de desempeño para el modelo de agrupamiento seleccionado. El criterio guía fue el índice de Dunn. 36

Tabla 4: Índices de desempeño para el modelo de agrupamiento sugerido por el enfoque de aglomerado jerárquico cortando para cinco conjuntos. 41

Tabla 5: Resultado de ejercicio de adaptación de los espectros promedios obtenidos mediante el algoritmo de agrupamiento *k-means* con la base de datos *KnowItAll*. 43

Tabla 6: Resultado del ejercicio de adaptación de los espectros promedios obtenidos mediante la estrategia de enfoque jerárquico confrontando la base de datos espectrales KnowItAll. . . . 46

Tabla 7: Distribución por estaciones de las muestras conforme al agrupamiento k-means. 51

Tabla 8: Distribución por estaciones de las muestras conforme al enfoque jerárquico aglomerativo. 51

Resumen

TÍTULO: Generación de un modelo numérico de agrupamiento (Clustering) para la identificación de microplásticos recolectados en el Parque Nacional Natural Corales de Profundidad y en el Parque Nacional Natural Los Corales del Rosario y San Bernardo, mediante espectroscopía infrarroja por transformada de Fourier con reflectancia total atenuada (FTIR-ATR). [1]

AUTORES: Diego Fernando Jaimes Castro[2], Isabel Cristina Prada Buitrago [3]

PALABRAS CLAVE: Microplásticos, Clustering, Aprendizaje no supervisado, Espectroscopía FTIR - ATR.

DESCRIPCIÓN:

Se proponen dos modelos de agrupamiento para elementos sólidos, con tamaño menor a 5 mm, en muestras de agua de mar, colectadas durante la temporada seca (Diciembre 2021 - Abril 2022) en doce estaciones de monitoreo localizadas en los Parques Nacionales Naturales Corales de Profundidad y Corales del Rosario y San Bernardo, a partir de sus espectros de absorbancia en el infrarrojo medio, con número de onda en el rango 4000 cm^{-1} – 500 cm^{-1}, usando FTIR-ATR.

A partir de 818 espectros FTIR-ATR, previamente normalizados y expresados en sus primeras siete (7) componentes principales (PCA por sus siglas en inglés, Principal Component Analysis), se implementó un modelo de cúmulos mediante el algoritmo *k-means*, el cual sugirió cinco (5) grupos, atendiendo a indicadores de validación intrínsecos (Dunn, Davies-Bouldin y Silueta). De manera similar, se desarrolló un dendrograma, que sintetiza el agrupamiento jerárquico. Los espectros promedios de los grupos sugeridos por los modelos se comparan con espectros de referencia de la base de datos comercial (KnowItAll, Bio-Rad/Wiley).

La comparación con la base de datos permitió identificar minerales típicos de la descomposición de los corales y algunos polímeros, a saber: polietileno, poliéster, polietileno tereftalato (PET), polipropileno. Dado que el tamaño de la muestra es menor a 5mm, se trata de microplásticos. Así, este estudio evidencia la presencia de estos elementos en estos sensibles ecosistemas.

[1] Trabajo de grado.

[2] Escuela de Física. Facultad de Ciencias. Jader Enrique Guerrero Bermúdez, Ph.D (GOTS, Director). Rafael Cabanzo Hernández, MSc (LEAM, Codirector).

[3] Escuela de Ingeniería Química. Facultad de ingenierías fisicoquímicas. Jader Enrique Guerrero Bermúdez, Ph.D (GOTS, Director). Rafael Cabanzo Hernández, MSc (LEAM, Codirector).

Abstract

TITLE: Numerical clustering model generation for the identification of microplastics collected in Parques Nacionales Naturales Corales de Profundidad y Corales del Rosario y San Bernardo, using Fourier Transform Infrared Spectroscopy with Attenuated Total Reflection (FTIR-ATR). [1]

AUTHORS: Diego Fernando Jaimes Castro[2], Isabel Cristina Prada Buitrago [3]

KEYWORDS: Microplastics, Clustering, Non supervised learning, FTIR - ATR spectroscopy.

DESCRIPTION:

Two clustering models are proposed for solid elements smaller than 5 mm in seawater samples collected during the dry season (December 2021 - April 2022) at twelve monitoring stations located in Parques Nacionales Naturales Corales de Profundidad y Corales del Rosario y San Bernardo, based on their mid-infrared absorbance spectra in the range of 4000 cm^{-1} – 500 cm^{-1}, using FTIR-ATR .

Using 818 pre-normalized FTIR-ATR spectra expressed in their first seven (7) principal components (PCA, Principal Component Analysis), a clustering model was implemented by means of *k-means* algorithm, which suggested five (5) clusters based on intrinsic validation indicators (Dunn, Davies-Bouldin, and Silhouette). Similarly, a dendrogram was developed to summarize the hierarchical clustering. The average spectra of the clusters suggested by the models were compared with reference spectra from the commercial database (KnowItAll, Bio-Rad/Wiley).

The comparison with the database allowed the identification of typical minerals from coral decomposition and some polymers, namely: polyethylene, polyester, polyethylene terephthalate (PET), and polypropylene. Since the sample size is smaller than 5mm, these are microplastics (MPs). Thus, this study provides evidence of the presence of these elements in mentioned ecosystems.

[1] Bachelor's final work.

[2] Escuela de Física. Facultad de Ciencias. Jader Enrique Guerrero Bermúdez, Ph.D (GOTS, Advisor). Rafael Cabanzo Hernández, MSc (LEAM, Co-advisor).

[3] Escuela de Ingeniería Química. Facultad de ingenierías fisicoquímicas. Jader Enrique Guerrero Bermúdez, Ph.D (GOTS, Advisor). Rafael Cabanzo Hernández, MSc (LEAM, Co-advisor).

Introducción

A partir de la segunda mitad del siglo XX el uso del plástico ha permanecido en constante crecimiento, al punto de ser inconcebible un mundo sin él, alcanzando una producción (global) de 359 millones de toneladas métricas en el año 2018 (Rocha-Santos et al., 2022; Oberbeckmann et al., 2014). Este crecimiento tan acelerado se vio potenciado por la introducción de los plásticos de un solo uso preferidos sobre los reutilizables en el mercado del embalaje (Geyer et al., 2017). Sin embargo, es más preocupante aún la cantidad de desechos plásticos que se generan debido a inadecuados manejos y políticas de disposición de estos residuos, por esto sólo un porcentaje muy pequeño se recicla, mientras la alarmante mayoría termina acumulándose en vertederos o siendo arrojados al medioambiente, perpetuándose así un proceso de contaminación imparable.

Se estima que en 2010 alrededor de 275 millones de toneladas métricas de desechos plásticos fueron generadas por comunidades costeras y que entre 4.8 y 12.7 millones de toneladas fueron introducidas al mar (Jambeck et al., 2015). Por su parte, los ríos conducen al mar entre 0.41 y 4 millones de toneladas que transportan desde tierra adentro (Garcés-Ordóñez et al., 2021). Los desechos plásticos ingresan a los cuerpos hídricos en forma de objetos completos, como empaques, botellas, bolsas y otros similares. Sin embargo, debido a su interacción con condiciones ambientales como la salinidad, la radiación solar y la oxidación, entre otros factores, estos desechos se degradan y fragmentan gradualmente hasta alcanzar dimensiones inferiores a 5 mm. A estas partículas se les denomina microplásticos. (Zhang et al., 2021).

Los microplásticos (MPs) constituyen una preocupación para comunidades científicas y organismos de sanidad por variadas razones: se generan permanentemente y tardan mucho en degradarse, ocasionando que se acumulen en la superficie de mares y océanos, en el suelo marino, en manglares, playas y otros ecosistemas costeros, amenazando la biodiversidad ya que por los diferentes tamaños, formas y colores de los fragmentos. Los organismos que habitan estos ecosistemas, desde zooplancton hasta grandes vertebrados, confunden estas piezas plásticas con alimento (Garcés-Ordóñez et al., 2021; Jung et al., 2018). Todo lo anterior empeora por las pequeñas dimensiones de los MPs que hacen de su erradicación un proceso difícil y costoso.

El camino más práctico para el control de este tipo de contaminación es la prevención, evitando que lleguen a cuerpos hídricos mediante un manejo riguroso de los desechos plásticos producto de actividades antropogénicas. Saber cuál es el polímero predominante en las muestras de microplásticos ayudará a establecer información acerca de los elementos de origen y la industria a la cual pertenecen, permitiendo enfocarse en establecer o rediseñar estrategias de reciclado y manejo de estos residuos plásticos. Más aún, se podría aspirar a impulsar nuevas tecnologías para la producción de polímeros cuyo proceso de degradación no constituya una

amenaza para ningún ser vivo (Jung et al., 2018).

Por ello los grupos de investigación LEAM, GOTS (adscritos a la Escuela de Física) y CEIAM (interdisciplinario) de la Universidad Industrial de Santander, buscan estimar y proporcionar datos de referencia sobre la contaminación de microplásticos en el marco del proyecto "Evaluación de la contaminación por microplásticos sobre la comunidad de zooplancton en el Parque Nacional Natural Corales de Profundidad y en el Parque Nacional Natural Los Corales del Rosario y de San Bernardo" con área comprendida entre: 9°43'16.59" -10°7'30.27" N (latitud), 75°47'16.25" – 76°17'41.09" O (longitud), frente a la costa de los departamentos colombianos de Bolívar y Sucre. Una buena gestión de disposición de desechos plásticos requiere conocer la cadena y ruta en los arrecifes, pastos marinos y manglares, donde constituyen elementos contaminantes y vehículo de otros tantos, amenazando la vida de los individuos que conforman estos hábitats (Acosta et al., 2019, 2018).

La técnica utilizada en el proceso de caracterización de las muestras colectadas en campo es la espectroscopía de infrarrojo por transformada de Fourier – reflexión total atenuada (ATR – FTIR) haciendo uso del espectrofotómetro Nicolet iS50 FT-IR. Se buscó agrupar las muestras tomadas en doce estaciones, durante la temporada seca (Diciembre - Abril, de acuerdo con el IDEAM, Instituto de Hidrología, Meteorología y Estudios Ambientales de Tierras), en los ecosistemas mencionados. Para ello se desarrollaron dos modelos para aglomerar: el modelo de k medias móviles, que en adelante llamaremos por su nombre técnico *k-means*, y un segundo modelo apoyado en un enfoque distinto: cúmulos jerárquicos que termina con la elaboración de un dendrograma. Una vez obtenida las diferentes asociaciones para las muestras, se les ubicó de acuerdo con las estaciones de muestreo y se propuso un ejercicio de adaptación (identificación de compuestos o especies) enfrentando los espectros promedio o representativos contra bases de datos reconocidas para infrarrojo (Primpke et al., 2018).

Justificación

Colombia es un país que, como pocos, tiene salida hacia dos importantes cuerpos hídricos, el Océano Pacífico y el Mar Caribe, con $928.660 km^2$ correspondientes a territorio marítimo. Su línea costera se extiende aproximadamente por $3200 km$, sobre el océano pacífico ($1576 km$) y el mar Caribe ($1642 km$), abarcando doce departamentos con poblaciones residentes en sus zonas costeras e insulares que suman cerca de 6'300.000 habitantes al cierre del año 2019. Cerca del 87 % corresponde a la región Caribe (Ministerio de Ambiente y desarrollo sostenible, 2019). Es decir, que las poblaciones en zonas costeras del país correspondieron al 12.6 % de la totalidad de la población nacional para el 2019, partiendo de una base de 50.187.406 habitantes en Colombia para ese mismo año, según datos del Banco Mundial (2022).

Es evidente la relación de proporcionalidad directa entre la densidad poblacional de una región, su desarrollo económico y un incremento en la generación de desechos (Geyer et al., 2017). La región Caribe de Colombia, con su extensa población residente, su alto movimiento turístico y, por consiguiente, con un relativamente alto desarrollo económico, es entonces de particular interés para las investigaciones de contaminación de sus aguas marinas por desechos plásticos. Más importante aun considerando que cerca del 79 % de los $1800 km^2$ de la superficie de corales que tiene Colombia está ubicada en el Caribe colombiano (Ministerio de Ambiente y desarrollo sostenible, 2021).

En el Caribe colombiano se encuentran el Parque Nacional Natural Corales de Profundidad (PNNCPR) y Parque Nacional Natural Los Corales del Rosario y de San Bernardo (PNNCRSB), dos áreas protegidas que hasta la fecha no han sido objeto de investigación de contaminación por microplásticos. Si bien Garcés-Ordóñez et al. (2021) realizaron un estudio para determinar la abundancia, distribución y características fisicoquímicas de MPs recolectados en aguas costeras superficiales del litoral pacífico y caribe colombiano, se identificó una oportunidad de aporte para la ampliación de información no solamente acerca de los MPs en sí, sino extendiendo literalmente las zonas de exploración desde aguas costeras a aguas mar adentro del Caribe colombiano.

1. Marco teórico

1.1 Los microplásticos en los océanos

Actualmente se ha incrementado la revisión teórica y experimental para comprender los mecanismos de degradación de los plásticos en estuarios, bahías y otros hábitats marinos en diversas partes del mundo (Nakano and Arakawa, 2022; Walsh et al., 2021; Tosin et al., 2012; Robin et al., 2020). El plástico en la superficie del mar, sometido a la intemperie y principalmente por la acción de la radiación solar (mayoritariamente Ultravioleta) se reduce a fragmentos, algunos de ellos tan pequeños, que pueden ser confundidos con alimentos por animales diminutos, cuyas comunidades constituyen la base de cadenas tróficas (Desforges et al., 2015; Andrady, 2011; Jiang et al., 2022; Wesch et al., 2016). Se suma a esto, la facilidad con que estos residuos, en forma de láminas, fibras, cuentas, etc, sirven de vehículo, al adherirse a su superficie otros elementos de reconocida toxicidad (vectores de xenobióticos) (Miri et al., 2022; Acosta et al., 2019). En estos términos, la persistencia y proliferación de residuos plásticos, de cualquier tamaño, constituye una amenaza real para la vida en los ecosistemas marinos.

La polución con plásticos, en especial los fragmentos menores a $5mm$, constituye una crisis global, desafío que requiere una respuesta conjunta para el monitoreo y el control de la acumulación y distribución de estos residuos sobre la superficie de la tierra (Bergmann et al., 2015; Zhang et al., 2022). Independiente de su origen, como material particulado (micro cuentas) que resulta de un proceso industrial (MPs primarios) o por desintegración inducida por condiciones externas (MPs secundarios), se trata de un problema en ascenso, a una tasa proporcional al descuido de la emisión de estos residuos, y con denuncias que datan desde principio de 1970 (Gago et al., 2018; Geyer et al., 2017; Gago et al., 2016).

El polímero sometido a la intemperie representa un desafío para su identificación. La idea es establecer su historia y ruta hacia el océano para controlarlo de manera más eficiente desde la fuente (Seghers et al., 2022). Así, el polímero prístino, del cual procede un fragmento catalogado como MP, eventualmente exhibe diferencias significativas en sus características fisicoquímicas. Es reconocido que los polímeros tienen una basta heterogeneidad, los más frecuentes en el ambiente marino, de acuerdo a la literatura, son los siguientes grupos: Polipropileno, poliestireno, polietileno, polietileno tereftalato (PET), cloruro de polivinilo (PVC), poliamidas (Chen et al., 2022; Zhu et al., 2019). En este trabajo es de especial interés la huella espectroscópica

en el infrarrojo medio, que sirve de característica para agruparlos de manera automática. Los modelos de agrupamiento son una estrategia básica cuando se carece de muestras con etiquetas. A partir de los conjuntos propuestos es posible consultar bases de datos (o incluso proponer nuevas referencias) (Primpke et al., 2018).

1.2 Espectroscopía infrarroja por transformada de Fourier - reflexión total atenuada. (FTIR - ATR)

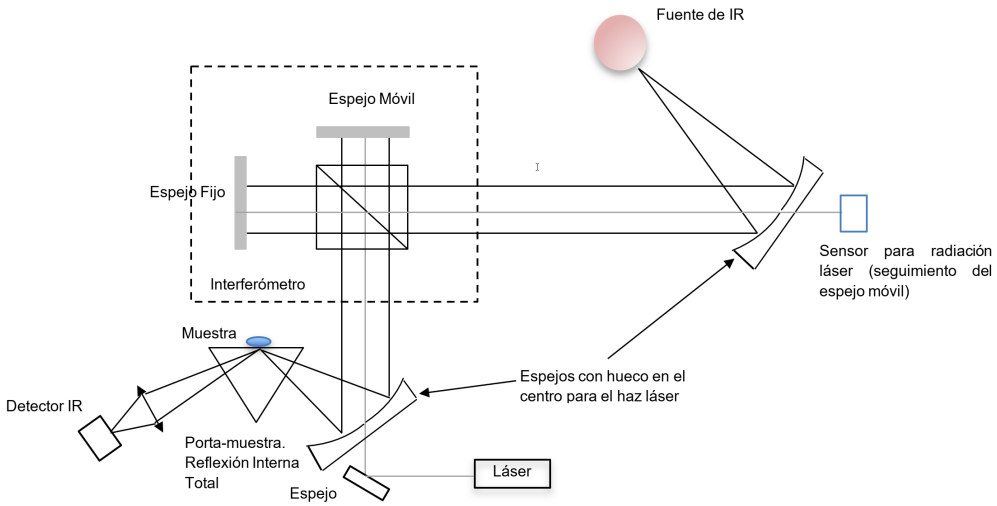
La espectroscopía infrarroja por transformada de Fourier (FTIR por sus siglas en inglés) y la espectroscopía Raman se encuentran entre las técnicas más notables para abordar el estudio de MPs debido a su huella vibracional (Dong et al., 2022; Weisser et al., 2022; Anger et al., 2018; Shim et al., 2017). En la literatura se propone la utilización de espectroscopía Raman para MPs $< 50\mu m$ y espectroscopía FTIR para MPs $> 50\mu m$ pues para medidas inferiores a los $20\mu m$, esta técnica presenta un porcentaje considerable de error debido a la difracción. Resalta que con la utilización de ambas técnicas de manera complementaria se puede lograr un barrido total de los tamaños de MPs comprendidos entre $1\mu m$ y $5000\mu m$ (Käppler et al., 2016). En este trabajo se utiliza la técnica de espectroscopía ATR - FTIR para la medición de espectros de MPs.

La espectroscopía infrarroja ha sido una técnica ampliamente utilizada desde su desarrollo por parte de William Weber Coblentz a comienzos del siglo XX, quien descubrió que utilizando una termopila era posible capturar el espectro infrarrojo (IR) que antes había sido elusivo a ser medido debido a la utilización de placas fotográficas para la visualización de espectros, y cuya sensibilidad se reducía a partir de los $650nm$, de tal forma que servían para espectros UV y visible pero tenían muy poca sensibilidad a longitudes de onda arriba del rojo (Kaur, 2021; Skoog, 2019). Inicialmente los montajes experimentales para espectroscopía infrarroja se basaron en la dispersión a partir del uso de rejillas, lo cual hacía la adquisición de espectros una tarea muy demandante de tiempo y esfuerzo con una baja relación de señal/ruido requiriendo equipos de gran tamaño para resultados precisos y procesos repetitivos de calibración.

Los avances tecnológicos de la segunda mitad del siglo XX permitieron el desarrollo de la espectroscopía FTIR implementando la extraordinaria precisión del conocido interferómetro de Michelson, que haciendo uso de espejos motorizados genera múltiples interferogramas, los cuales se procesan aplicando una transformada de Fourier para obtener el espectro IR resultante, aumentando la razón señal a ruido (Abidi, 2022). Este aumento propició la masificación de la espectroscopía FTIR, incluyendo algunas variantes como la FTIR-ATR que aprovecha la atenuación de la reflexión interna total. La Figura 1, muestra un esquema de un espectrómetro típico de este tipo de espectroscopía infrarroja (Skoog, 2019). En la reflexión interna total, debido a la diferencia de índices de refracción en la interfase, existe un ángulo de incidencia crítico para el cual no existe componente transmitida, sino una onda que decae exponencialmente en amplitud y se encuentra

confinada a la interfase, se conoce como onda evanescente y constituye el principio de funcionamiento del ATR - FTIR al interactuar con la muestra (Kaur, 2021; Blum and John, 2012).

Figura 1.
Esquema de montaje óptico de un espectrómetro *ATR - FTIR*.



Nota. Tomado de Skoog (2019).

En el espectrómetro *ATR - FTIR* utilizado, la reflexión interna total atenuada del haz modulado por el interferómetro de Michelson, sucede en un prisma de diamante y el medio absorbente con el cual interactúa la onda evanescente es la muestra de MP. La disipación de energía electromagnética en esta interacción genera una atenuación en la onda incidente. Esta característica del espectrómetro *ATR - FTIR* aporta ventajas importantes para la obtención de espectros a partir de muestras en estado sólido o líquido (Milosevic, 2012).

La importancia del segmento IR del espectro electromagnético en la espectroscopía radica en que las transiciones entre estados de energía vibracional en átomos o moléculas dadas por $3N - 6$ grados de libertad (donde N representa el número de átomos) ocurren en este rango de energía, con lo cual se pueden explorar composiciones moleculares evaluando las bandas de absorbancia características de grupos funcionales conocidos y a su vez tener información sobre su conformación y estructura molecular (Blum and John, 2012).

1.3 Fundamentos del tratamiento de datos: Reducción de dimensionalidad mediante PCA (Análisis de componentes principales).

Con la reducción de la dimensionalidad se busca una representación más compacta de los datos de entrada, conservando la información relevante que cada uno de ellos contiene, disminuyendo así la complejidad

computacional, mejorando la escalabilidad del modelo y de forma general su desempeño (Tharwat et al., 2017). El método de reducción de dimensionalidad utilizado en este proyecto, conocido como *PCA* (por sus siglas en inglés: *principal component analysis*) es un enfoque de extracción de características en el cual se tiene inicialmente una serie de vectores de datos con d dimensiones, proyectados en un espacio dimensional l ($d > l$), mediante una base de vectores ortogonales conocidos como componentes principales (Forsyth, 2019).

En la aplicación del método PCA se busca que esta transformación lineal maximice la varianza de la proyección de los datos de entrada sobre la componente principal ω_i, lo cual podemos expresar como:

$$\begin{aligned} Var\left(\omega_i^\top x\right) &= E\left[\left(\omega_i^\top x - \omega_i^\top \mu\right)\left(\omega_i^\top x - \omega_i^\top \mu\right)\right] \\ &= E\left[\omega_i^\top \left(x-\mu\right)\left(x-\mu\right)^\top \omega_i\right] \\ Var\left(\omega_i^\top x\right) &= \omega_i^\top \sigma_x^2 \omega_i \end{aligned} \tag{1}$$

En donde σ_x^2 representa la covarianza, x los datos de entrada y /mu... . Para maximizar la varianza con la componente principal ω_i se plantea lo obtenido como un problema de Lagrange sujeto a la ligadura $\omega_1^T \omega_1 = 1$.

$$\begin{aligned} \mathcal{L}(x_i, x_j, \lambda) &= f(x_i, x_j) + \lambda g(x_i, x_j) \\ &= \omega_1^\top \sigma_x^2 \omega_1 - \lambda\left(\omega_1^\top \omega_1 - 1\right) \qquad (2) \\ \frac{\partial \mathcal{L}}{\partial \omega_i} = 0 &\Rightarrow 2\sigma_x^2 \omega_1 - 2\lambda\omega_1 = 0 \\ \sigma_x^2 \omega_1 &= \lambda \omega_1 \qquad (3) \end{aligned}$$

Lo expresado en la ecuación (3) se cumple siempre que la componente principal ω_1 sea un *vector propio* de la covarianza σ_x^2 y λ su respectivo *valor propio*, de esta forma para maximizar la varianza expresada en la ecuación (1) se reduce a cero el segundo término en la ecuación (2) y se toma el *vector propio* cuyo *valor propio* sea mayor, con lo cual obtenemos que la primera componente principal ω_1 para la cual se resolvió la ecuación (3) es la dirección en el espacio sobre la cual las proyecciones tienen una mayor varianza (Alpaydin, 2014). De manera equivalente es posible calcular las demás componentes principales que conforman todo el espacio d dimensional. Conforme se avanza en las componentes, decrece su varianza y en consecuencia es posible eliminar las últimas componentes principales con baja varianza obteniendo una representación muy aproximada de la serie de datos de entrada con una dimensionalidad reducida.

1.4 Modelos de agrupamiento.

Para los distintos algoritmos de agrupamiento es posible identificar dos enfoques principales que se diferencian en la forma de entrenar el modelo y también en las condiciones que deben cumplir los datos de entrada. Por una parte, tenemos los algoritmos de *aprendizaje supervisado* los cuales se caracterizan por su naturaleza clasificatoria, en donde el modelo es entrenado a partir de unos datos de entrada y una salida etiquetada (característica clasificatoria), de tal forma que éste adquiera habilidad predictiva y clasifique de manera precisa nuevos datos de entrada no conocidos dentro de las etiquetas preestablecidas en el entrenamiento (Johnston et al., 2019; Bonaccorso, 2018).

Por otra parte, los algoritmos de *aprendizaje no supervisado* carecen de un criterio que establezca previamente las etiquetas, permitiendo al modelo explorar correlaciones y patrones de distribución en los datos de entrada y generar agrupamientos en función de sus hallazgos. Este será el enfoque de aprendizaje utilizado en el presente trabajo, debido a la naturaleza de los fragmentos colectados, los cuales, en principio, no deberían ajustarse a las señales espectrales de polímeros prístinos (o cualquier otra especie distinta a polímeros) y por lo tanto no es adecuado etiquetarlos a priori (Zhou, 2021; Celebi, 2015; Alpaydin, 2014).

1.4.1 Modelos de aprendizaje no supervisado.

En el aprendizaje no supervisado se entregan datos de entrada no etiquetados al modelo. El cual se desarrolla a partir de los patrones o estructuras subyacentes que puedan ser detectados, relacionando y agrupando entradas con atributos similares entre sí. De esta forma, los modelos de agrupamiento son una valiosa herramienta de exploración de datos que proporciona información sobre la distribución de éstos, facilitando la detección autónoma de grupos de características similares, así como de valores atípicos fuera de estos grupos, los cuales pueden ser de interés especial para aplicaciones en reconocimiento de patrones, procesamiento de imágenes, entre otros (Han et al., 2011).

Los algoritmos basados en este método forman una cantidad *preestablecida* de cúmulos de los datos de entrada a partir de la distancia entre atributos y el valor representativo que actúa como centroide. Algunos de estos métodos son *k-means* (k-medias móviles) y *k-medians* (k-medianas móviles) de los cuales abordaremos *k-means*, utilizado en el desarrollo de este proyecto.

1.4.2 *k-means.*

Debido a la relativa simplicidad de su implementación para datos de tipo numérico y su garantía de convergencia, *k-means* es un algoritmo muy utilizado. Para su inicialización debe elegirse un número k que define el número de grupos a formar a partir de los datos de entrada, tomando como valor representativo la media de cada grupo y asignando un valor entrante al centroide más cercano, de manera que dicho centroide es recalculado para cada iteración al considerar los elementos que se añaden al grupo, buscando converger de acuerdo a una tolerancia establecida (Reddy and Vinzamuri, 2021; Zhou, 2021).

Se recomienda fuertemente un proceso de pre-aglomerados; esto contribuye a lograr una convergencia más rápida y grupos más compactos. De cualquier modo, es necesario evaluar el desempeño del agrupamiento mediante índices apropiados, que se mencionarán más adelante. Un Algoritmo 1 para implementar *k-means* se presenta a continuación.

Algoritmo 1 Implementación de *k-means.*

Entrada: Conjunto de n datos de entrada $x = \{x_1, x_2, \ldots, x_n\}$, número k de grupos y valor de tolerancia.
Salida: Centroides finales de los grupos $c_1, c_2, ..., c_k$.

1: Se toman k centroides iniciales (aleatorios o sugeridos por algún proceso de pre-agrupamiento).
2: **mientras** la variación en posición del centroide sea mayor que tolerancia **hacer**
3: Se asigna cada uno de los datos de entrada al aglomerado más cercano midiendo las distancias entre estos y los centroides presentes.
4: Se calcula la nueva posición de cada centroide.
5: **fin mientras**
6: **devolver** Centroides finales de los grupos $c_1, c_2, ..., c_k$.

1.4.3 Agrupamiento jerárquico.

También se ha implementado un análisis exploratorio de datos mediante un modelo de agrupamiento jerárquico, el cual considera (en principio) cada espectro en el espacio de componentes principales como un grupo individual, que a través de un proceso iterativo se unirán de dos en dos, mediante un criterio de similaridad, hasta formar un único aglomerado, generando así un árbol cuya representación gráfica se conoce como dendrograma (Aggarwal and Reddy, 2013; Han et al., 2011).

Para explicar un algoritmo que genere grupos jerárquicos es necesario profundizar en la forma de medir la similaridad entre pares de conjuntos y así poder agruparlos; para ello se requiere establecer dos importantes características que son la *métrica* y el *criterio de similaridad (linkage)*. La métrica proporciona

información sobre la distancia entre cualquier par de elementos con la misma dimensionalidad, por lo cual cada métrica tendrá un efecto en la estructura misma del agrupamiento jerárquico. El *linkage* o vínculo, por otra parte, establece la regla que describe cómo fusionar los elementos que muestran la mayor afinidad (Dougherty, 2012). El vínculo que resulta más común es el *promedio*. También son frecuentes los valores máximos o mínimos de las muestras que van a fusionarse. Se propone un Algoritmo 2 aglomerativo con relaciones de jerarquía.

Si bien la complejidad temporal del algoritmo depende en parte del tipo de relación o vínculo utilizado, todos ellos habitualmente requieren calcular una matriz completa de distancias.

Algoritmo 2 Implementación de agrupamiento aglomerativo con relaciones de jerarquía.

Entrada: n datos de entrada $x = \{x_1, x_2, \ldots, x_n\}$, una métrica $D(\cdot, \cdot)$ y un criterio de similaridad.

Salida: Un agrupamiento conformado por un único elemento, que incluye a todos los datos de entrada.

1: Se inicializa con n grupos, tomando cada dato de entrada o espectro como grupo unitario (un solo elemento).
2: **mientras** se obtiene un único aglomerado que contenga todos los datos. **hacer**
3: Según la métrica elegida (distancia euclidiana, distancia Manhattan, etc), se calcula la distancia entre cada uno de los grupos.
4: Se encuentra el par de grupos con mayor semejanza, según el criterio de similaridad elegido (promedio, máximo, etc.)
5: Se fusionan ambos grupos.
6: **fin mientras**
7: **devolver** Un agrupamiento conformado por un único elemento, que incluye a todos los datos de entrada.
8: A partir de las uniones encontradas se grafica un dendrograma y luego se corta a la altura de un k a elegir.

1.5 Índices de validación interna

Con los índices de validación interna se busca evaluar la calidad de la solución encontrada por un algoritmo de agrupamiento, basándose únicamente en la información presente en los datos de entrada, de tal forma que en ausencia de etiquetas proporciona información determinante para encontrar qué modelo funciona mejor para esos datos y también el número óptimo de aglomerados (Maulik and Bandyopadhyay, 2002).

Un agrupamiento ideal busca que los elementos dentro de un mismo conjunto tengan la máxima similaridad posible, y a su vez, al comparar dos elementos en diferentes conjuntos tengan la mayor diferencia o distancia posible; es decir, un grupo *compacto tiene una máxima distancia intragrupo pequeña y una mínima separación intergrupal grande*(Liu et al., 2010; Halkidi et al., 2001). En este sentido, existen múltiples índices o medidas de validación interna que consideran los criterios anteriores implementados de diferentes formas, con lo cual la evaluación de la calidad del agrupamiento ocurre en función de parámetros distintos . En consecuencia, distintos índices pueden resultar más apropiados para evaluar diferentes tipos de datos y diferentes algoritmos

de agrupamiento (Ansari et al., 2015; Halkidi et al., 2001).

Teniendo en cuenta lo anterior, se abordan a continuación los índices de validación utilizados en el presente trabajo, escogidos para funcionar de manera complementaria, observando que sus tendencias estén conformes con su definición.

1.5.1 Índice de Dunn

El índice de Dunn utiliza la relación entre la mínima distancia entre medias de aglomerados distintos (separación intergrupal) y la mayor separación de un elemento a la media del grupo al que pertenece (compactibilidad intragrupo). Para una propuesta de agrupamiento con m grupos, el índice de Dunn (DI_m) de este modelo puede calcularse mediante:

$$DI_m = \frac{min_{1\leq i<j\leq m}\delta(C_i, C_j)}{max_{1\leq i<j\leq m}\Delta_k} \tag{4}$$

En la ecuación anterior, $\delta(C_i, C_j)$ es la distancia entre las medias de los conjuntos C_i y C_j. Por otro lado, Δ_k, es distancia del k-ésimo elemento a la media del grupo al que pertenece. Para diferentes modelos, se recomienda el seleccionar el que exhiba el coeficiente de Dunn más alto, garantizando así grupos compactos y separados (Zhou, 2021; Ansari et al., 2015; Dunn, 1973).

1.5.2 Índice de Davies-Bouldin

Dado una propuesta de agrupamiento, el índice Davies-Bouldin de este modelo, se calcula considerando la razón $R_{i,j}$ para la pareja conformada por los conjuntos C_i y C_j. El numerador de esta fracción es la suma de los promedios de la distancia de los elementos a los centroides de su grupo. $<d_i>$ y $<d_j>$, son respectivamente los promedios a los centroides de C_i y C_j. El denominador, es la distancia entre las medias del par de grupos considerados, $d_{Cen}(i, j)$.

$$R_{i,j} = \frac{<d_i> + <d_j>}{d_{Cen}(i, j)}$$

El promedio de los valores más altos $R_{i,j}$ representado en la ecuación (5) constituye el índice de Davies-Bouldin (Liu et al., 2010).

$$D_i = max_{i \neq j} R_{i,j}$$
$$DB = \frac{1}{N} \sum_{i=1}^{n} D_i \tag{5}$$

Para garantizar grupos compactos, con alta separabilidad es recomendable elegir el modelo con el número de grupos que presente un valor pequeño en el índice de Davies-Bouldin (Halkidi et al., 2001; Davies and Bouldin, 1979).

1.5.3 Coeficiente de siluetas simplificado

El valor silueta simplificado se calcula para cada elemento i, que conforma el conjunto de datos, como la diferencia de la distancia del elemento al centroide más próximo $b(i)$ y al centroide al cual pertenece $a(i)$, partida por el mayor valor entre ellos, esto es:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{6}$$

Es decir, el índice de siluetas mide que tan bien un elemento se ajusta a su grupo en comparación con los otros conjuntos, de tal forma que el valor del índice varía entre -1 y 1, donde un valor cercano a 1 indica que el objeto está bien ajustado a su aglomerado y separado de los otros, mientras que un valor cercano a -1 indica que el objeto no se encuentra bien agrupado (Lenssen and Schubert, 2022; Kaufman and Rousseeuw, 2005).

Ahora bien, para efectos de reportar, qué tan bien agrupados, se encuentran todos los elementos, se utiliza el coeficiente silueta, que se relaciona con la media de los valores calculados mediante la ecuación (6).

2. Metodología

2.1 Obtención de muestras

Las muestras se recolectaron en doce estaciones ubicadas dentro de las áreas marinas protegidas (AMP) Parque Nacional Natural Corales de Profundidad (PNN CPR) y del Parque Nacional Natural Los Corales del Rosario y de San Bernardo (PNN CRSB) localizadas frente a las costas de los departamentos de Bolívar y Sucre 9°43'16. 591" - 10° 7'30. 277" N (latitud) 75°47'16.254" - 76°17'41.091" O (longitud). El muestreo se realizó del 17 al 22 marzo de 2022, correspondiente a la temporada seca.

Para la recolección de las muestras de la columna de agua, se utilizó una red minibongo de $30cm$ de diámetro equipada con mallas de $200\mu m$ y $500\mu m$ con vaso colector de PVC. Se realizó un recorrido oblicuo durante 5 minutos a $5{,}5km/h$ de velocidad. El volumen de agua filtrada se determinó utilizando un flujómetro Hydrobios (Modelo No. 438115) sin motor de reversa situado en el centro de la boca de la red de $200\mu m$.

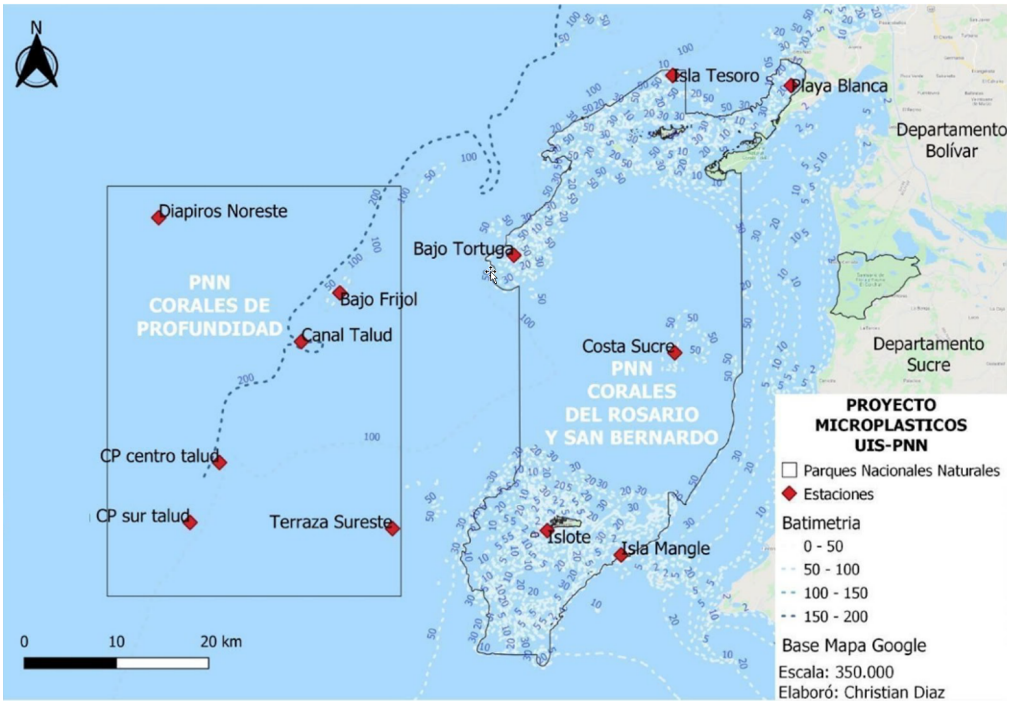
En la recolección de sólidos superficiales, se siguió el protocolo de Kovač et al. (2016) con modificaciones , se utilizó una red manta de $65cm$ de ancho y $30cm$ de alto, con una malla de $300\mu m$ con vaso colector metálico, en un recorrido superficial con el 60 % del área de la red sumergida. Se realizó un recorrido de 30 minutos a $5{,}5km/h$ de velocidad. El volumen de agua filtrada se determinó utilizando un flujómetro General Oceanic (Modelo No. $2030R$) situado en la parte sumergida de la boca de la red. Las muestras superficiales se fijaron con etanol al 70 % para su posterior análisis, y las muestras de la columna de agua fueron fijadas en formaldehído buferizado al 4 %.

Las estaciones ubicadas en el PNN Corales del Rosario y San Bernardo corresponden a puntos cercanos a la costa, en tanto que las estaciones dentro del PNN Corales de Profundidad corresponden a puntos mar adentro, distribución que permitirá establecer una correlación entre el tipo de MP recolectados y la ubicación de las estaciones respecto a su cercanía a la costa. En la Tabla 1, se presentan las estaciones en orden de recolecta, con su nombre correspondiente y el área protegida en el que se encuentra ubicada cada una de ellas; complementando la tabla se encuentra un mapa de la zona en el que se visualiza la distribución geográfica de estos puntos en las áreas señaladas (Figura 2).

Tabla 1.
Estaciones de muestreo para la exploración de microplásticos en los Parques Nacionales Naturales (PNN) Corales del Rosario y San Bernardo (CRSB) y Corales de Profundidad (CPR).

No.	Nombre de la estación	PNN
1	Playa Blanca	CRSB
2	Canal Talud	CPR
3	Isla Tesoro	CRSB
4	Bajo Frijol	CPR
5	Islote	CRSB
6	Terraza Sureste	CPR
7	Bajo Tortuga	CRSB
8	Formación CP Sur Talud	CPR
9	Costa Sucre	CRSB
10	Formación CP Centro Talud	CPR
11	Isla Mangle	CRSB
12	Diapiros Noreste	CPR

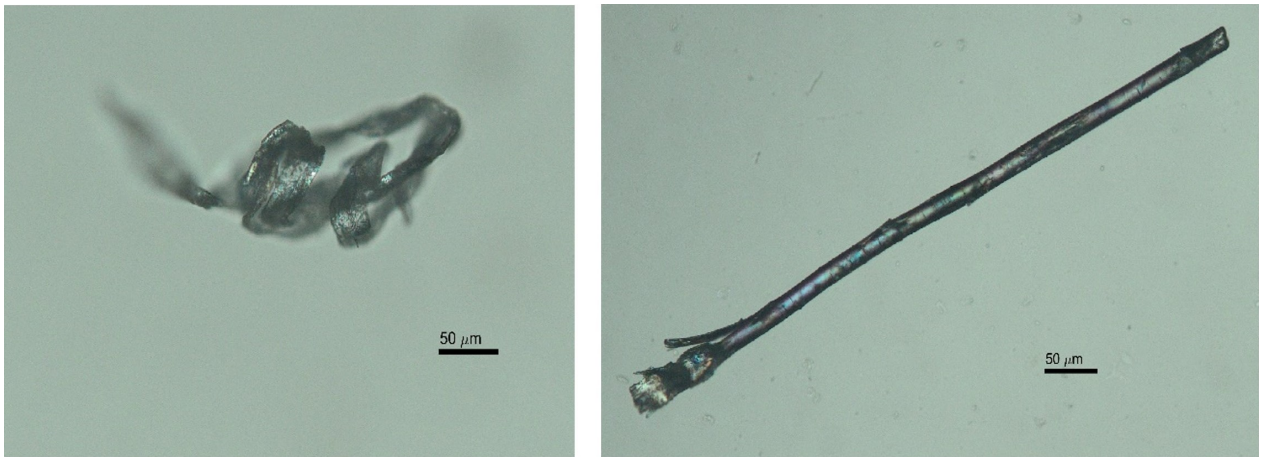
Figura 2.
Ubicación geográfica de las estaciones de medición en los parques nacionales naturales Los Corales del Rosario y San Bernardo, y Corales de Profundidad.



2.2 Entrega de muestras

Para el desarrollo de la sección experimental de este trabajo de investigación, las muestras de agua tomadas en cada estación se filtraron y por inspección, mediante el uso de un estereoscopio, se separaron manualmente todos aquellos sólidos cuyo diámetro estuviera por debajo de los $5mm$, criterio para clasificarlo en la escala micro. Las muestras fueron entregadas por lotes correspondientes a cada estación de muestreo, cada grupo de muestras clasificadas según su forma (filamentos, fragmentos, espumas, pellets o películas); según el color (rojo, azul, blanco, negro u otros) y la malla en la que fueron encontrados. Las muestras secas y separadas, fueron entregadas en medio de dos portaobjetos, asegurados con cinta adhesiva y rotuladas adecuadamente, indicando: el código del proyecto, la salida de campo en la que fueron recolectadas, la estación, la malla, la forma de la muestra sólida y su color. Un par de imágenes de las muestras típicas colectadas se presentan en la Figura 3.

Figura 3.
Las muestras más comunes fueron fibras laminares y cilíndricas como las mostradas a la izquierda y derecha de la figura, respectivamente.



2.3 Protocolo de muestreo

Como se indicó anteriormente, las muestras sólidas venían separadas por lotes correspondientes a la estación donde fue recolectada la muestra de agua. Esta misma clasificación fue respetada mientras se pasaban las muestras por el espectrómetro y para su conservación posterior a su análisis en el equipo.

Previo a la toma de cada espectro se limpiaba el prisma de diamante y la punta del torquímetro (pieza metálica que prensa la muestra contra el cristal) con un trozo de gaza de celulosa, humedecida con alcohol isopropílico para retirar de las superficies cualquier sustancia o material ajeno que pudiera contaminar

el registro. Seguidamente al análisis de la muestra en el equipo, se guardó el archivo electrónico del espectro generado conservando el rótulo que identificaba la muestra al llegar al laboratorio, añadiendo un número al final para identificar cada muestra en particular de acuerdo con el orden en que fueron analizadas en el espectrómetro.

2.3.1 Parámetros instrumentales

Los espectros fueron registrados en el infrarrojo medio $4000cm^{-1}$ - $500cm^{-1}$ mediante el *Thermo Scientific Nicolet iS50 FT-IR Spectrometer* con accesorio prisma de diamante *iD5 ATR, Thermo Fisher Scientific Madison, WI, USA*. La señal fue muestreada en 7468 puntos, con intervalo aproximado entre datos 0.48 cm-1. La velocidad del espejo móvil en el Michelson se estableció en $0{,}4747cm/s$ y 128 barridos por muestra. Para el posterior análisis se elimina el intervalo $2200cm^{-1}$ - $1900cm^{-1}$, excluyendo la banda anómala del índice de refracción del diamante.

Con estos parámetros se logra para los espectros de absorbancia un promedio en la razón señal a ruido, $<S/N> = 27{,}40$. Esta se calculó tomando el valor de la absorbancia (para todas las muestras) en el intervalo $1880cm^{-1}$ - $1831cm^{-1}$ (aproximadamente 100 puntos). Debe anotarse que los espectros son relativamente constantes en este intervalo. Antes de archivar el espectro, se realizó corrección de línea base mediante el software OMNIC del equipo. En la Tabla 2 se resumen los parámetros instrumentales para el registro de los espectros.

Tabla 2.
Los espectros de las muestras colectadas fueron registrados con los siguientes parámetros instrumentales.

Elemento	**Descripción**
Equipo	Thermo Scientific Nicolet iS50 FTIR
Modo	Absorbancia
Rango de adquisición	$4000cm^{-1}$ - $500cm^{-1}$
Número de barridos	128
Resolución	$4cm^{-1}$
Velocidad espejo móvil	$0{,}474cm/s$

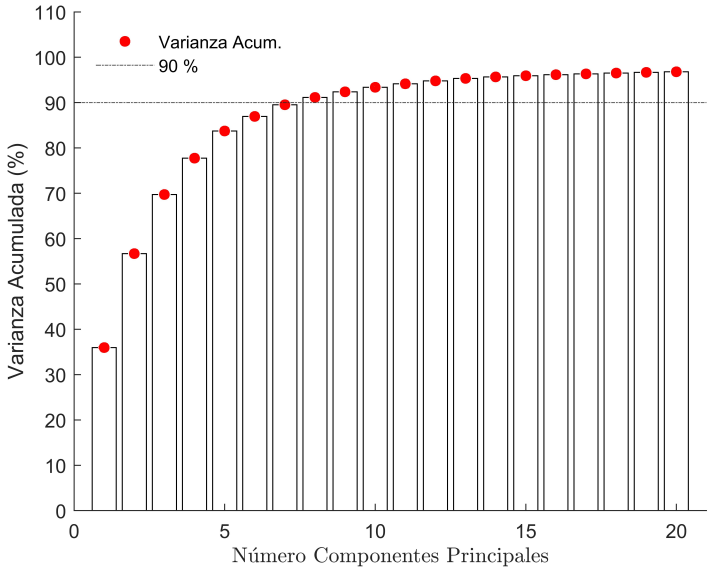
3. Análisis y discusión de resultados

3.1 Dimensionalidad obtenida mediante PCA

Antes de cualquier procesamiento numérico, los espectros de absorbancia fueron sometidos a una corrección de línea base y normalizados por el área bajo la curva espectral. Con el propósito de proyectar este conjunto de espectros en un espacio vectorial de dimensionalidad reducida, se estiman las componentes principales mediante el algoritmo NIPALS (*Nonlinear Iterative Partial Least Squares*) para lo cual se requiere definir previamente la cantidad de componentes principales a encontrar. La determinación de la dimensionalidad óptima es un problema comúnmente abordado en la literatura y pese a que existen varias reglas o criterios para escoger la cantidad de componentes, éstas no significan una solución única pues dicho valor dependerá del conjunto de datos de entrada y el problema que con ellos se busque resolver (Otto, 2016).

Con el objetivo de determinar la dimensionalidad que mejor represente el conjunto de datos de entrada, se considera el comportamiento de la varianza acumulada versus el número de componentes.

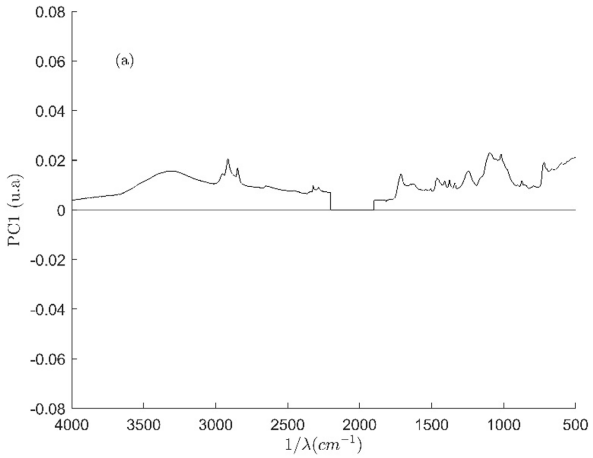
Figura 4.
Representación de la varianza acumulada en función del número de componentes principales. La señal se explica por encima del 90 % al considerar más de siete componentes principales.



La varianza acumulada para las primeras 20 componentes se muestra en la Figura 4, se puede observar que a partir de la séptima componente se explica más del 90 % de la señal. Adoptamos este porcentaje como un umbral aceptable, considerando además que no se genera una ganancia significativa en la varianza acumulada (y en la varianza explicada) a partir de este número de componentes. Esta base vectorial reduce en cerca de tres órdenes de magnitud la dimensionalidad asociada a cada espectro. De esta manera se disminuye la complejidad computacional, se elimina información redundante y se gana interpretabilidad en los datos. A manera de ejemplo, las primera siete componentes principales se muestran en la Figura 5(a) hasta (g).

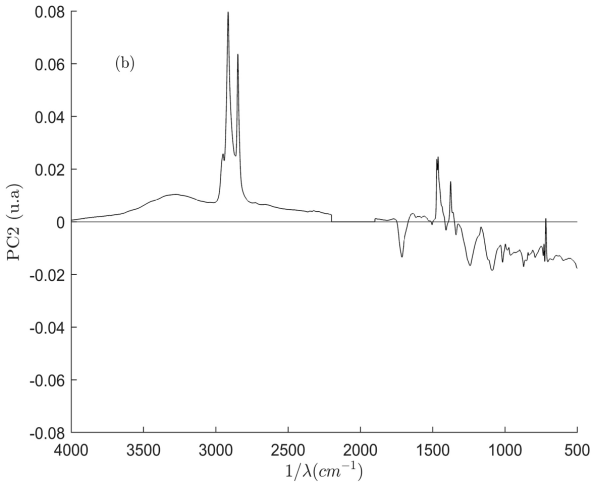
Figura 5.
Las siete primeras componentes principales conforman una base vectorial, de esta manera cada espectro se expresa mediante los siete coeficientes que ponderan estos vectores propios.

Figura 5(a)



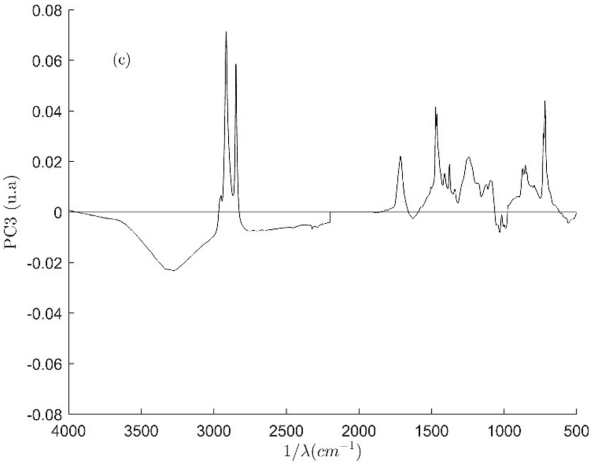
Nota. Primera componente principal (PC1).

Figura 5(b)



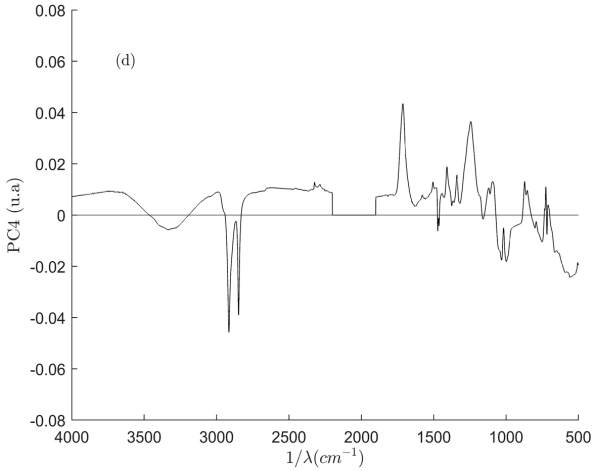
Nota. Segunda componente principal (PC2).

Figura 5(c)



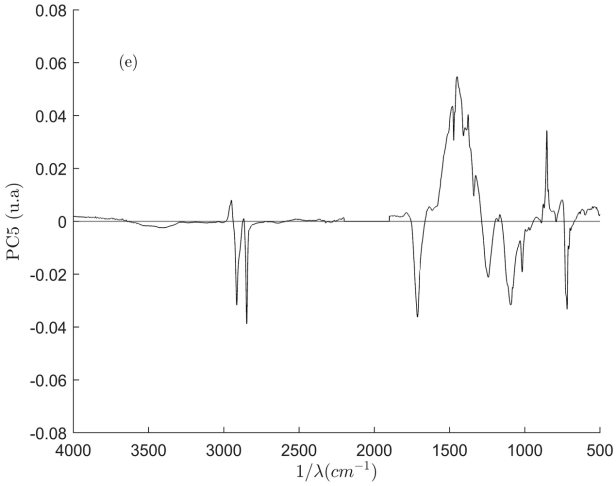
Nota. Tercera componente principal (PC3).

Figura 5(d)



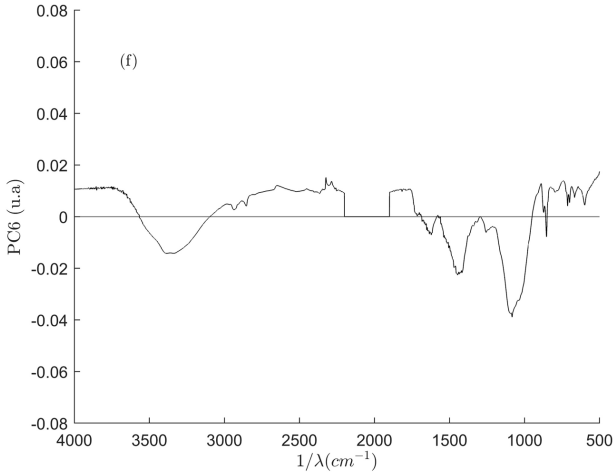
Nota. Cuarta componente principal (PC4).

Figura 5(e)



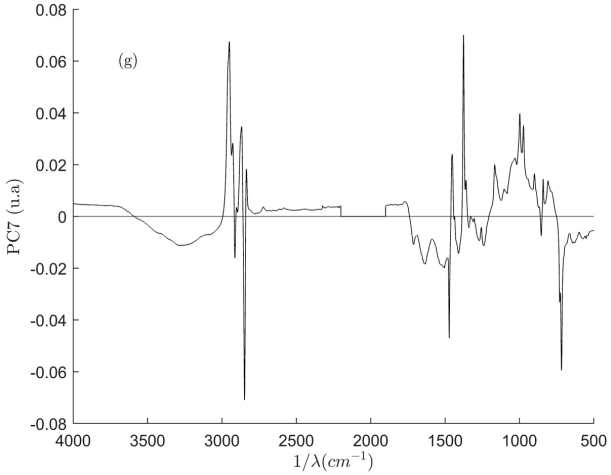
Nota. Quinta componente principal (PC5).

Figura 5(f)



Nota. Sexta componente principal (PC6).

Figura 5(g)



Nota. Séptima componente principal (PC7).

Cada PC constituye un espectro de absorbancia 'sintético', donde las regiones destacadas (positivas o negativas) dan información de las bandas de absorbancia (transmisión) asociables a las especies químicas que se quieren estudiar.

En el primer componente principal, Figura 5(a), se observa que las magnitudes, en unidades arbitrarias, son todas positivas, y es posible identificar claramente varias bandas: $3600cm^{-1} - 3000cm^{-1}$ que se puede asignar al estiramiento del grupo funcional OH; una banda delgada con dos picos intensos en el rango entre $3000cm^{-1} - 2800cm^{-1}$ que corresponde al modo vibracional de estiramiento de los grupos CH_2 o CH_3 alifáticos (Socrates, 2001).

En la región de la huella espectral característica de los microplásticos, por debajo de los $1800cm^{-1}$, se tiene una banda intensa y delgada cerca de $1700cm^{-1}$ atribuida al estiramiento del grupo funcional $C = O$. Una banda conformada por varios picos en la región $1470cm^{-1} - 1300cm^{-1}$ que tiene origen en las vibraciones del tipo flexión (tijereteo) del grupo funcional CH alifático, así como flexiones del grupo CH_2 (Mukherjee et al., 2018; Socrates, 2001).

La banda de mayor intensidad en la región entre $1150cm^{-1} - 900cm^{-1}$, se asigna al estiramiento del grupo funcional $C - O$, así como algunos compuestos plastificadores basados en fosfatos ($1000cm^{-1}$) y grupo funcional $O - C - O$ ($1100cm^{-1}$). Este conjunto de bandas representa el de mayor ocurrencia en el conjunto de datos iniciales.

En la segunda componente principal, Figura 5(b), se observan valores negativos sobre la zona principal de huella espectral de los microplásticos. Las regiones de mayor relevancia en PC2 serán la banda de estiramiento de CH_2, CH_3 alifáticos ($3000cm^{-1} - 2800cm^{-1}$) y de flexión del grupo CH_2 ($1450cm^{-1} - 1350cm^{-1}$). Este grupo de bandas son características del polipropileno.

La tercera componente principal, Figura 5(c), presenta un comportamiento interesante, principalmente priorizando el estiramiento CH_2 y zonas puntales de la región de huella espectral característica mientras que se da peso negativo al estiramiento OH. Esto evidencia que en PC3 el algoritmo busca separar los espectros que cuenten con solo algunas de las características más comunes detectadas. Desde PC4 hasta PC7 (Figura 5(d) - 5(g)), ocurre un comportamiento similar, entregando pesos negativos a las bandas de estiramiento CH y OH mientras que se prioriza los valores ubicados dentro de la zona de huella espectral característica.

3.2 Pretratamiento del conjunto de datos para *k-means*

Es bien conocido que la estrategia *k-means* requiere establecer previamente el número de grupos. Para determinar esta cantidad se utilizan índices o coeficientes, basados en la consistencia de los cúmulos, como los definidos en los fundamentos teóricos de esta monografía: el índice Dunn, el índice Davies-Bouldin y el coeficiente Silueta. Estas figuras de mérito se calcularon considerando todas las muestras disponibles (un total de 818), con una estrategia de pre-aglomerado utilizando la similaridad coseno, descrita a continuación.

Dado un número de componentes principales, se elige de manera aleatoria una muestra, que constituye la primera 'cabeza de grupo'. Esta muestra se compara con todas las restantes y se conforma un cúmulo con todos los espectros que exhiben una similaridad superior a un umbral dado (en la práctica se trabajó

con umbrales mayores al 79 % de similaridad e inferiores al 95 %). A continuación, se selecciona la siguiente muestra 'cabeza de grupo' y se realiza el mismo proceso hasta agotar todas las muestras. Es posible que queden muestras que no logren agruparse, formándose con ella un aglomerado residual. Este pre-agrupamiento constituye la entrada al algoritmo *k-means*. Este procedimiento se repitió de manera aleatoria cincuenta veces y, para cada intento, se calcularon los índices mencionados.

El resultado de esta estrategia que intenta determinar a priori el número de grupos se muestra en la Figura 6, Figura 7 y Figura 8 para el coeficiente de Dunn, el índice de Davies-Bouldin (DB) y el coeficiente de silueta, respectivamente. En todas las figuras, el literal **(a)** corresponde al pre-agrupamiento con siete componentes principales, los literales (e) e (i) se utilizan para once y quince componentes. Las componentes restantes se pueden consultar en el anexo (A)-(C). Los umbrales de similaridad se consideraron en el rango del 79 % al 95 %.

Figura 6.
Comportamiento del índice de Dunn para diferentes propuestas de modelo de agrupamiento, en función del número de grupos y número de componentes principales. En **(a)**, **(e)** e **(i)**, se consideran siete (7), once (11) y quince (15) componentes principales.

Figura 6(a)

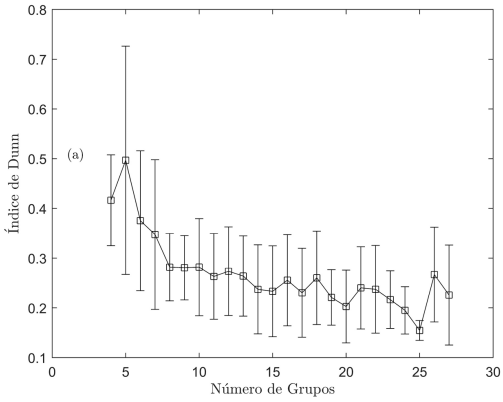


Figura 6(e)

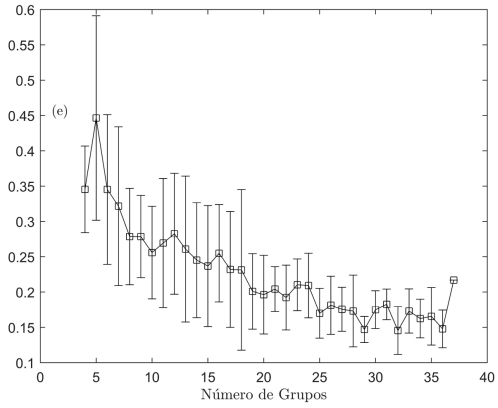


Figura 6(i)

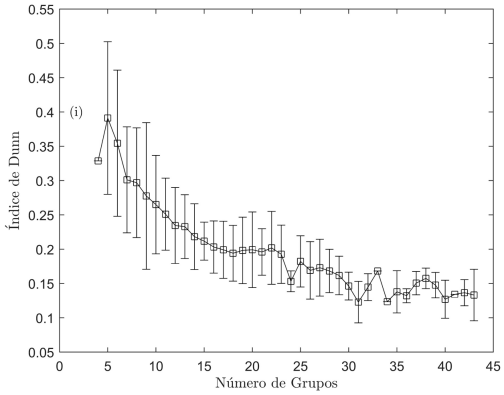


Figura 7.
Característica del índice de Davies-Bouldin para diferentes propuestas de modelo de agrupamiento. En **(a)**, **(e)** e **(i)**, se consideran siete (7), once (11) y quince (15) componentes principales. Nótese su tendencia contraria al coeficiente de Dunn, a medida que aumenta el número de grupos.

Figura 7(a)

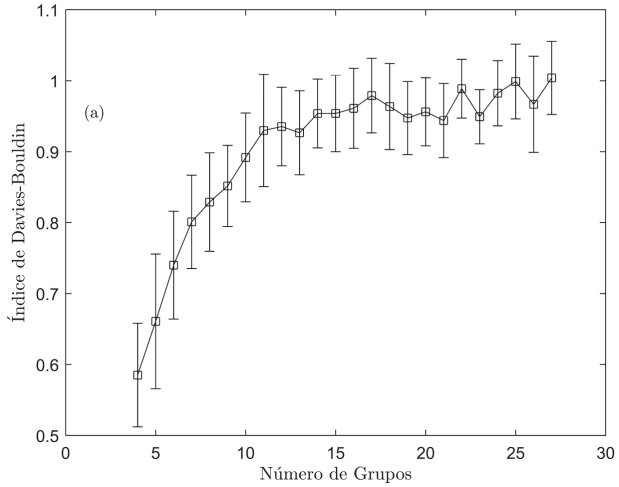


Figura 7(e)

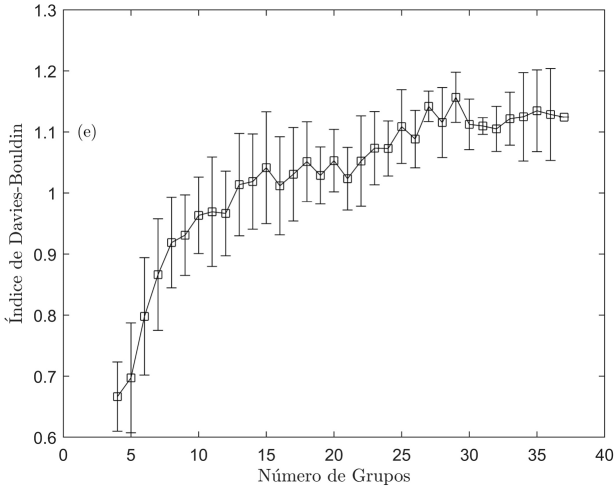


Figura 7(i)

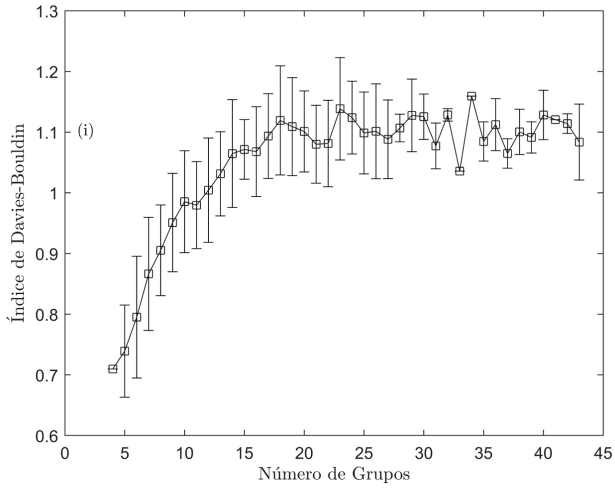


Figura 8.
Curva de coeficiente silueta versus número de grupos. En **(a)**, **(e)** e **(i)**, se consideran siete (7), once (11) y quince (15) componentes principales.

Figura 8(a)

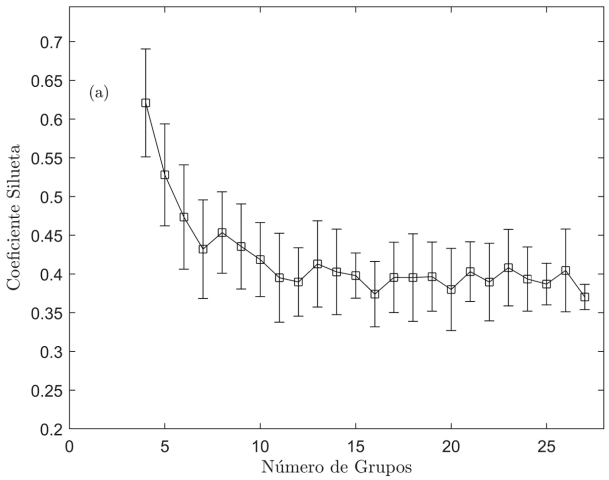


Figura 8(e)

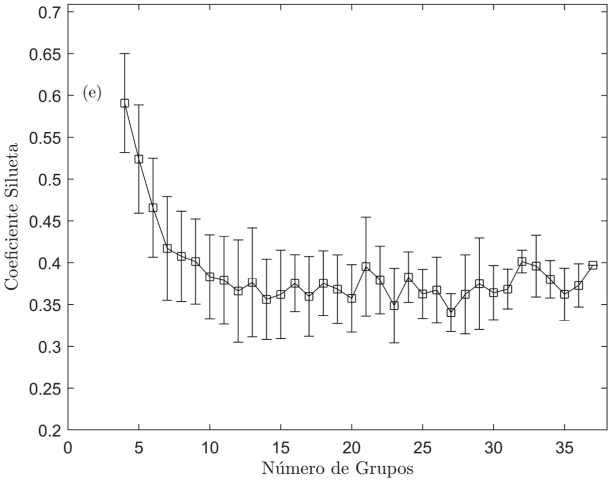
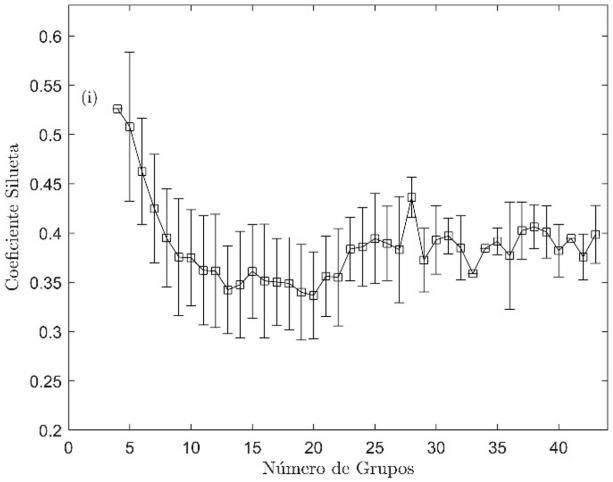


Figura 8(i)



Las características de la Figura 6, que describen el índice Dunn, muestran un comportamiento similar para cada una de las componentes. Conforme a la definición adoptada para el índice de Dunn, el número de grupos que se debe seleccionar es el correspondiente a su valor más alto, en este caso cinco (5) cúmulos. Por otro lado, se puede apreciar que conforme aumenta el número de componentes principales (y también el umbral de similaridad) se incrementa el número de aglomerados alcanzando un valor máximo próximo a los 45 (grupos).

Análogamente, el índice de Davies-Bouldin (DB) sugiere cuatro grupos (Figura 7). Un número de aglomerados compactos y de gran separación exhibe el coeficiente de DB más pequeño. Como en el caso anterior, para el índice de Dunn, las características del índice DB, al cambiar el número de componentes y el umbral de similaridad, presentan un comportamiento semejante. Finalmente, las curvas del coeficiente silueta (Figura 8) corroboran las sugerencias dadas por los índices anteriores, esto es, número relativamente pequeño de grupos (estrictamente 4 cúmulos). Una partición con coeficiente silueta alto da señales de grupos compactos y separados. Al igual que los índices anteriores, su comportamiento es bastante similar para cada número de componentes principales, umbral de similaridad y número máximo de aglomerados.

Así, ¿Qué cantidad de grupos seleccionar?, para el número de componentes y porcentaje de similaridad establecidos, los índices sugieren, desde su definición, pocos grupos. El índice de Dunn sugiere cinco grupos, los índices DB y Silueta cuatro. De acuerdo con este resultado se tomará como cinco el número de grupos para desarrollar el modelo de agrupamiento de los espectros registrados.

3.3 Resultados del agrupamiento *k-means*

Ajustado a los índices tratados en la sección anterior, se toman cinco (5) grupos como entrada al algoritmo *k-means* y una tolerancia para la medida de los desplazamientos de los centroides (con métrica la distancia euclidiana) de 10^{-3} unidades arbitrarias (u.a). Un código en Matlab (R2019b) del algoritmo *k-means* implementado por los autores de esta monografía se anexa a este documento (anexo D).

La partición de cinco grupos que se tomó como entrada al algoritmo tiene base vectorial conformada por 7 componentes principales y exhibe similaridad coseno superior al 85 %. Esta es una forma de correlación entre dos señales. Sus índices de desempeño se muestran en la Tabla 3. Para completar la información de los índices, se presenta la silueta en la Figura 9.

Tabla 3.
Índices de desempeño para el modelo de agrupamiento seleccionado. El criterio guía fue el índice de Dunn.

Índice	**Valor**
Dunn	0.89
Davies - Bouldin	0.58
Silueta	0.58

Figura 9.
Valor silueta para un modelo de cinco grupos, atendiendo al mejor índice de Dunn.

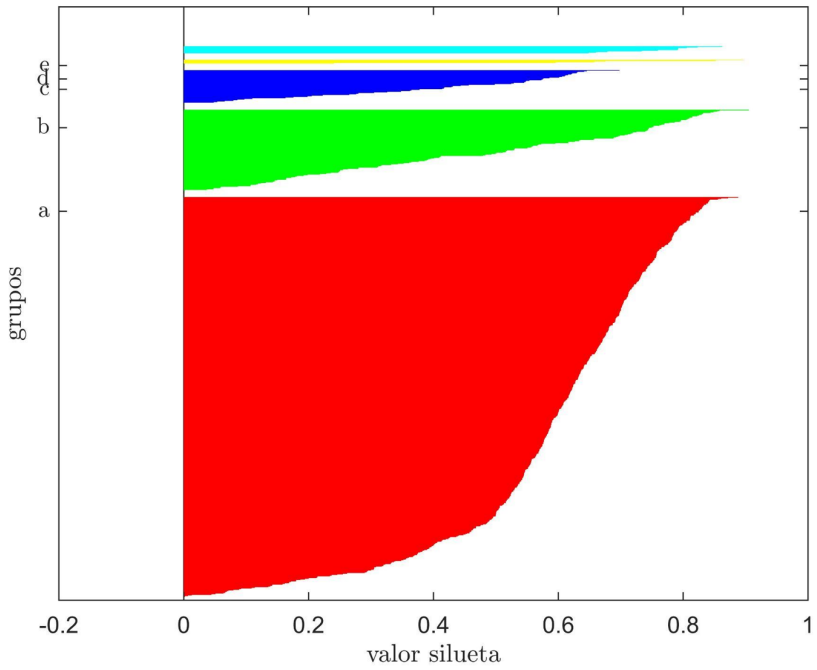


Figura 10.
La propuesta de cinco grupos sugiere los espectros promedio mostrados desde (a) hasta (e). Estos espectros promedio son instrumentos para explorar la presencia de MPs en las estaciones de los PNN considerados.

Figura 10(a)

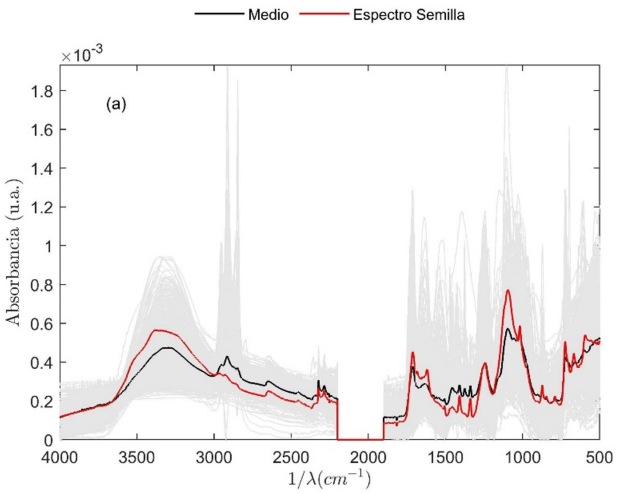


Figura 10(b)

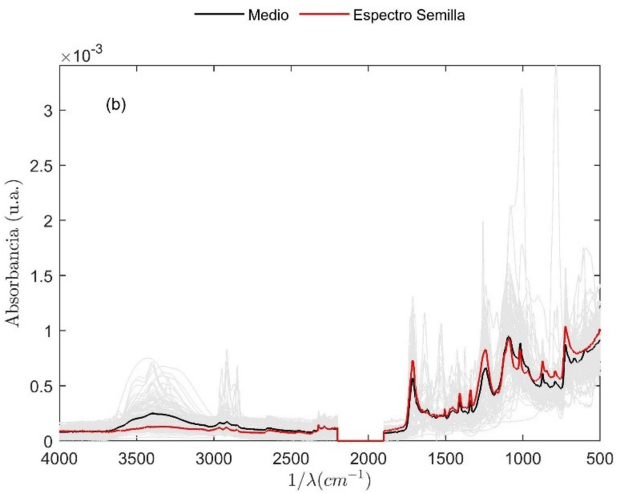


Figura 10(c)

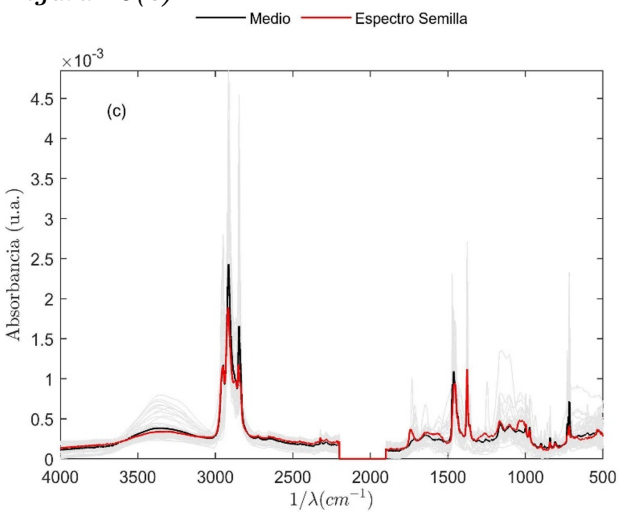


Figura 10(d)

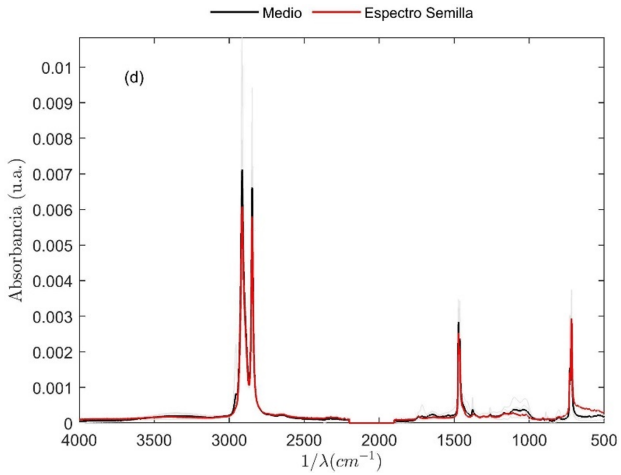
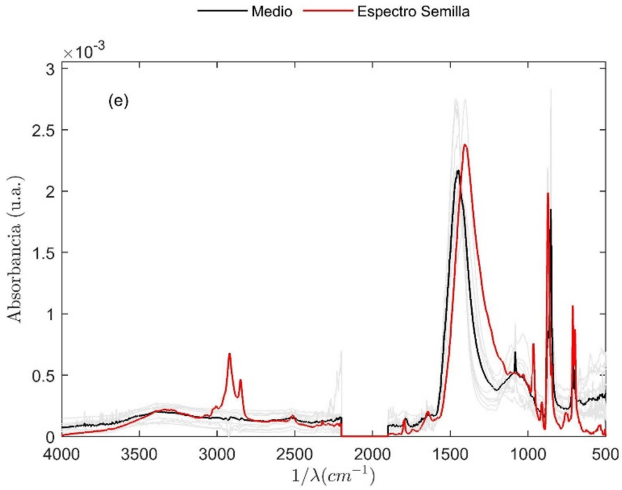


Figura 10(e)



La Figura 10(a)-**(e)** muestra los cinco conjuntos sugeridos, indicando en tono de gris los espectros de absorbancia de cada grupo. Se resalta en negrilla el espectro promedio para efecto de búsqueda e identificación de especies en las bases de datos. En rojo se muestra el espectro semilla o «cabeza de grupo», para indicar que tanto se desplaza la señal por la acción del algoritmo *k-means*.

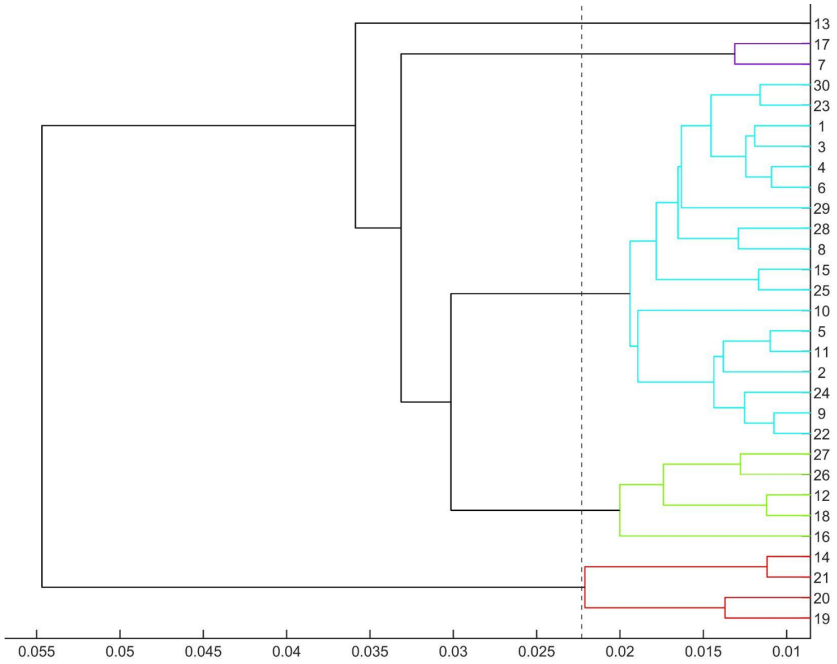
3.4 Resultados del agrupamiento por enfoque jerárquico: dendrograma

El enfoque jerárquico se implementó utilizando funciones del entorno Matlab R2019b. La entrada para crear los aglomerados son los espectros proyectados en las primeras siete componentes principales. Inicialmente, la función *linkage* retorna un arreglo (también llamado árbol) de tres columnas y $N-1$ filas, donde N es

el número total de muestras (818). Las dos primeras columnas relacionan el número de orden o etiqueta de los espectros más próximos (entre sí) y la tercera columna la distancia entre ellos. En este caso, la métrica para determinar la distancia entre espectros es la distancia euclidiana. El vínculo, fusión o *linkage* se realiza mediante el promedio (*average*) de los espectros más próximos.

La manera más efectiva de exhibir los resultados del agrupamiento jerárquico es mediante un **dendrograma**, es decir, una representación pictórica del arreglo o árbol conseguido mediante la función *linkage*. La función dendrogram genera, por defecto, 30 grupos iniciales cuando el número de muestras supera esta cantidad. Posteriormente, se vinculan o unen las 'hojas' más próximas con 'ramas' en forma de U (invertida o rotada), conformando así nodos, que en el caso de los dendrogramas aglomerativos se van reduciendo hasta conformar un único vínculo. La Figura 11 muestra el dendrograma que resulta con los espectros de absorbancia tratados.

Figura 11.
El dendrograma es la síntesis del enfoque aglomerativo. La línea de trazo permite discriminar cinco grupos (cantidad que sugieren los índices de validación interna). Debe observarse la similaridad existente entre la envergadura de las ramas en el dendrograma y el perfil de siluetas (Figura 9).



Para efectos de comparar con los resultados que se obtuvieron con el algoritmo *k-means*, el dendrograma se corta a una distancia cercana a 0,023($u.a$), ver Figura 11, obteniendo un número de cinco grupos de espectros de absorbancia (número sugeridos por los índices de validación interna). Los grupos de espectros se pueden apreciar en la Figura 12(a)-(e). En negrilla se presenta el espectro promedio del conjunto correspondiente.

Figura 12.
Para efecto de comparar con la estrategia *k-means* se corta el dendrograma de manera que resulten cinco grupos. En negrilla se presenta el efecto promedio.

Figura 12(a)

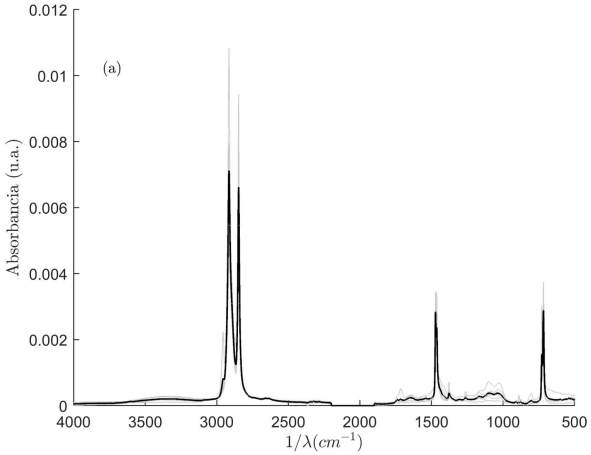


Figura 12(b)

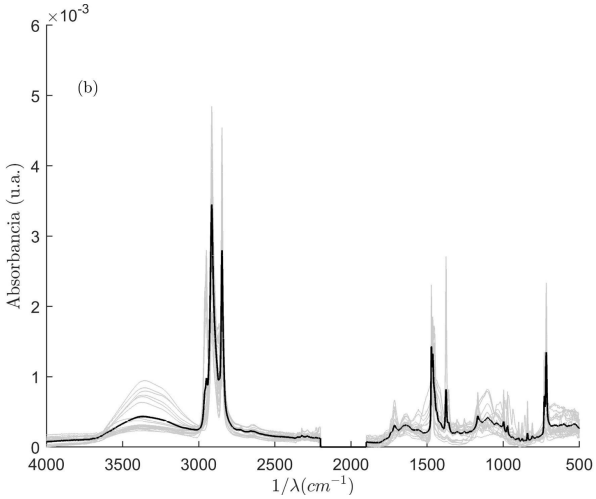


Figura 12(c)

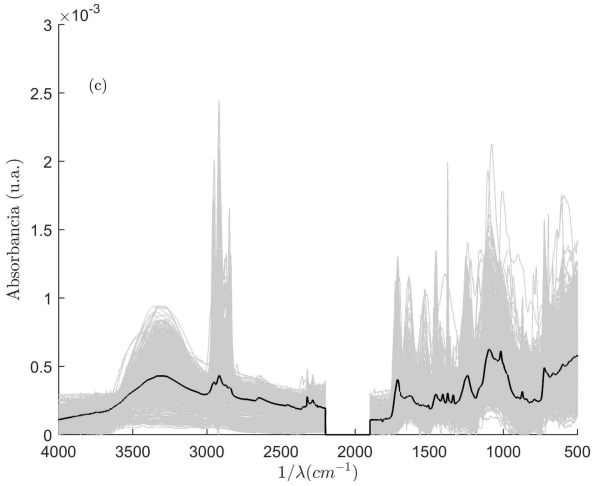


Figura 12(d)

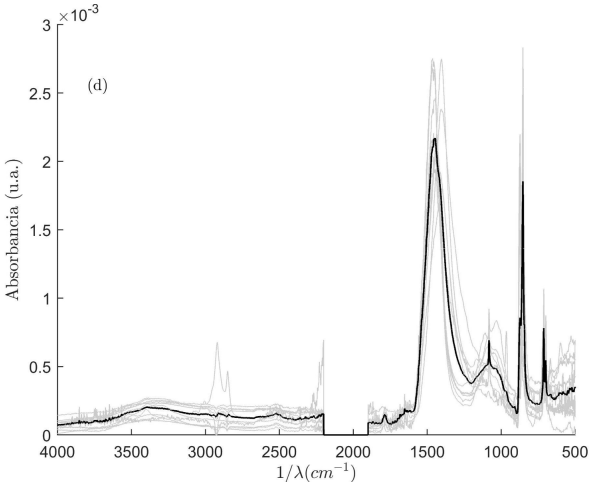
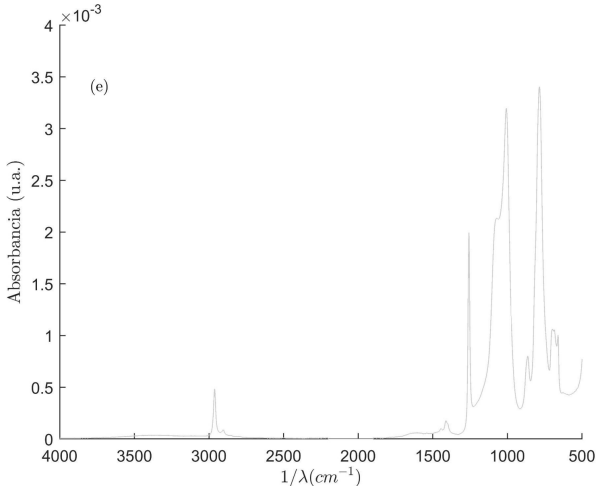


Figura 12(e)



A diferencia de los índices de la Tabla 3, que responden a una búsqueda exhaustiva para optimizar el agrupamiento, en la Tabla 4, se trata de información (no hay un proceso de optimización).

Tabla 4.
Índices de desempeño para el modelo de agrupamiento sugerido por el enfoque de aglomerado jerárquico cortando para cinco conjuntos.

Índice	Valor
Dunn	0.56
Davies - Bouldin	0.85
Silueta	0.45

3.5 Confrontación con base de datos

La base de datos espectral de referencia utilizada fue *KnowItAll* de *Bio-Rad/Wiley*, mediante la cual se establecieron las coincidencias entre espectros, atendiendo al HQI (por sus siglas en inglés *Hit Quality Index*), indicador que resulta ser proporcional a la correlación del espectro de interés con los espectros de la base. Esta búsqueda sugiere una lista de muchas especies, ordenadas en función del HQI. Con esta relación se explora la presencia de polímeros (polipropileno, poliestireno, polietileno, tereftalato de polietileno (PET), cloruro de polivinilo (PVC), poliamidas) o de especies características del ambiente marino. En particular, dado que la huella espectral de los MPs se modifica, por la acción de la intemperie, respecto a los polímeros prístinos (de los cuales se conjetura que proceden), es necesario completar la búsqueda de especies, iniciada con el HQI, mediante la adaptación de picos de absorbancia. Cabe resaltar que la región $2750cm^{-1} - 1850cm^{-1}$ se considera de baja especificidad y baja varianza, razón por la cual, en la búsqueda de MPs, no se tiene en cuenta al confrontar con la base de datos (Renner et al., 2019).

3.5.1 Grupos sugeridos por *k-means*

Cuando se confronta el espectro promedio del primer grupo obtenido mediante el algoritmo *k-means*, Figura 13(a), su correlación más alta con un espectro de referencia presenta un HQI de 63.23, en el cual se aprecian interesantes coincidencias en los picos de absorbancia. Se muestra el espectro representado en la base de datos con etiqueta *HPX 371* que corresponde a una mezcla en proporción 2:1 de poliéster y fibra textil de algodón. Un ejercicio similar se desarrolla para los otros cuatro grupos sugeridos por la estrategia *k-means*, Figura 13(b)-(e). La Tabla 5, muestra la especie con la adaptación más apropiada para cada uno de los grupos. En las Figura 13(a)-(e) y Figura 14(a)-(e), se coloca el espectro de referencia de la base de datos espectrales invertido para facilitar la comparación.

Profundizando en el análisis sobre esta adaptación entre los grupos sugeridos por el algoritmo *k-means* y la base de datos *KnowItAll* es importante considerar:

En el espectro del primer grupo, los picos de absorbancia coinciden en regiones importantes asociadas con la huella espectral del poliéster, entre ellas: $2920cm^{-1} - 2850cm^{-1}$ asociados a los estiramientos simétricos y asimétricos (respectivamente) del enlace $C - H$ en grupos funcionales metilo ($-CH_3$); el pico en $1720cm^{-1}$ que se asocia a la vibración de estiramiento del enlace carbonilo ($-C = O$) presente en los enlaces carbono - oxígeno de los grupos funcionales éster; el pico ubicado en $1460cm^{-1}$ que corresponde a la vibración de flexión del enlace $-CH_2-$ (grupo funcional metileno) presente en sus enlaces $C - H$; el pico en $1245cm^{-1}$ que se relaciona con la vibración de estiramiento del enlace $C - O - C$ presente en el grupo éster, y finalmente el pico ubicado en $1100cm^{-1}$ el cual se asocia a la vibración de estiramiento del enlace $C - O$ presente en la estructura del éster, siendo estos algunos de los picos de mayor intensidad en los cuales se encontró coincidencia, sin embargo se observa correspondencia en casi todos los demás picos de intensidad observables por encima de los $600cm^{-1}$, ver Figura 13(a). (Socrates, 2001).

La mejor correspondencia encontrada a partir del segundo agrupamiento obtenido, con un índice HQI de 68.37, corresponde al Tereftalato de polietileno (PET), como se puede observar en la Figura 13(b). Este presenta una coincidencia importante respecto al caso anterior tanto en los picos espectrales descritos como en las intensidades de absorbancia, ubicados en la región de la huella espectral de los MPs comprendida entre $1850cm^{-1}$ y $700cm^{-1}$, lo cual sugiere que sus estructuras químicas son similares (Renner et al., 2017).

La principal diferencia entre ambos espectros promedio (Figura 13(a) y Figura 13(b)) está en la intensidad de un pico de absorbancia ancho que se distribuye a lo largo de la región comprendida entre $3600cm^{-1}$ y $3200cm^{-1}$ el cual se relaciona con vibración de estiramiento en grupos funcionales hidroxilo ($-OH$). En el primer agrupamiento, este pico presenta una mayor intensidad y puede asociarse a la fibra natural del algodón que contiene celulosa, presente en la mezcla de poliéster con fibra textil. Por otra parte, el segundo agrupamiento evidencia una baja intensidad en la banda que caracteriza al grupo hidroxilo, asociándose al PET de tipo resina.

En la Figura 13(c) se aprecia el tercer grupo, que sugiere coincidencia con polipropileno (75 %) (PP), polietileno (19 %) (PE) y ácido poliacrílico (6 %) (PAA), con un índice HQI de 81.68. Se analiza en el espectro sus regiones representativas, encontrando tres picos de absorbancia en la banda $2960cm^{-1} - 2850cm^{-1}$ correspondientes al pico de mayor intensidad en $2920cm^{-1}$ asociado a la vibración de estiramiento asimétrico del enlace $C - H$ en el grupo metilo ($-CH_3$), un pico cerca de $2850cm^{-1}$ que corresponde al estiramiento asimétrico del enlace $C - H$ en grupos metileno ($-CH_2-$) y el tercero en $2950cm^{-1}$ que representa el

estiramiento simétrico del enlace $C-H$ en grupos metilo ($-CH_3$). También está presente el doblete metilo que corresponde a su flexión simétrica y asimétrica respectivamente en los picos $1370cm^{-1}$ y $1465cm^{-1}$ (Socrates, 2001).

El espectro del polietileno (PE) es el que presenta la mejor coincidencia con el espectro promedio del cuarto grupo, exhibe una correlación, representada por un índice HQI, de 92.60 y en cuyo espectro promedio, Figura 13(d), se puede identificar fácilmente el ya conocido doble pico asociado a vibraciones de estiramiento simétrico y asimétrico en los grupos metilo ($-CH_3$) ubicados en $2915cm^{-1}$ y $2850cm^{-1}$, respectivamente. Así también, el pico en $1470cm^{-1}$ que se relaciona con la flexión del enlace $C-H$ en grupos metilo ($-CH_3$) y el pico en $720cm^{-1}$ que describe la flexión del enlace $C-H$ en grupos metileno ($-CH_2-$).

En el quinto conjunto la mejor correlación corresponde al espectro del mineral aragonito con un índice HQI de 82,63. Este mineral es una forma cristalina del carbonato de calcio ($CaCO_3$) y puede indicar importantes procesos en el ecosistema marino relacionados con la presencia de corales y otros organismos calcificadores. Cabe resaltar que casi el 70 % de elementos que conforman este agrupamiento, provienen de la estación de recolección Canal talud, que se ubicó sobre una región con abundancia significativa de corales de profundidad.

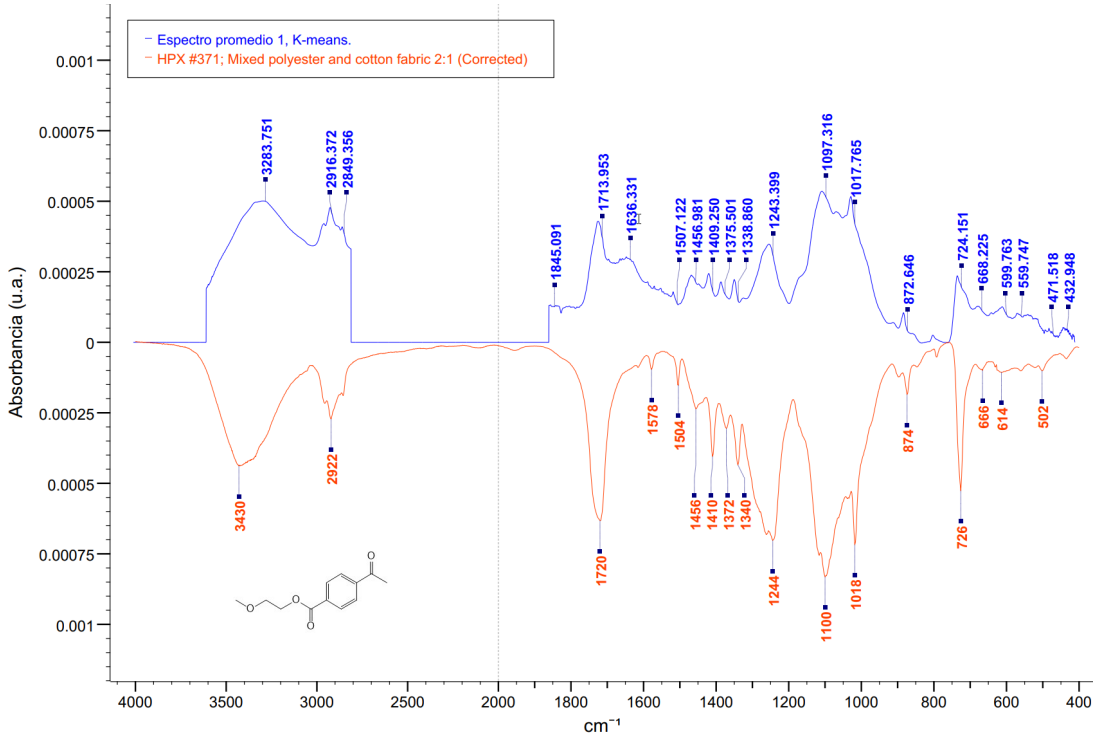
Se observa un corrimiento de algunos picos de absorbancia entre el espectro promedio y el sugerido por la base de datos. En particular, los cuales corresponden a un pico de gran intensidad en $1470cm^{-1}$ y $1450cm^{-1}$ que se asocian a la flexión del enlace CO_3 en el aragonito, un pico cerca de $850cm^{-1}$ que se relaciona con la vibración de estiramiento asimétrico del enlace CO_3. Igualmente el pico en $700cm^{-1}$ que corresponde a la flexión del enlace CO_3.

Tabla 5.
Resultado de ejercicio de adaptación de los espectros promedios obtenidos mediante el algoritmo de agrupamiento k-means con la base de datos KnowItAll.

	Coincidencia	**Espectro de referencia Base de datos**	**HQI**
Grupo a	Poliéster - Algodón	HPX 371	63.23
Grupo b	Tereftalato de Polietileno (PET)	BWX 144	68.37
Grupo c	Polipropileno, Polietileno, Ácido Poliacrílico	WX 2311	81.68
Grupo d	Polietileno	BWX 335	92.60
Grupo e	Aragonito	MNX 73	86.47

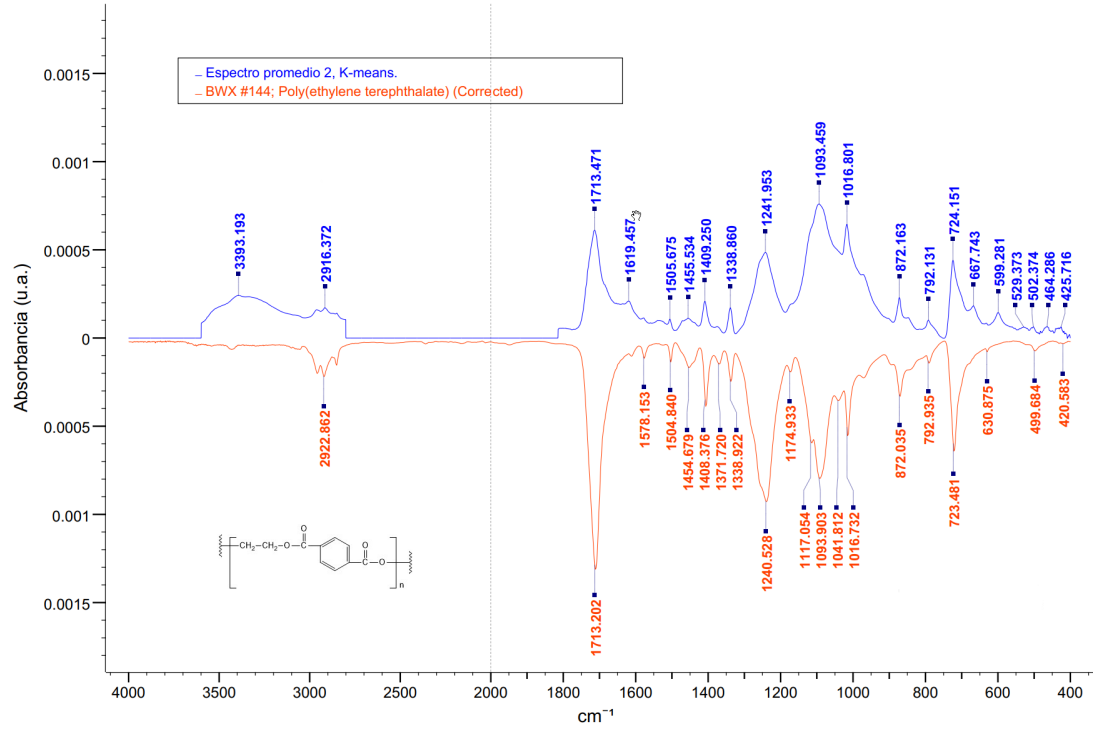
Figura 13.
Adaptación de los espectros promedio calculados por la estrategia *k-means* con espectros sugeridos por la base de datos *KnowItAll*.

Figura 13(a)



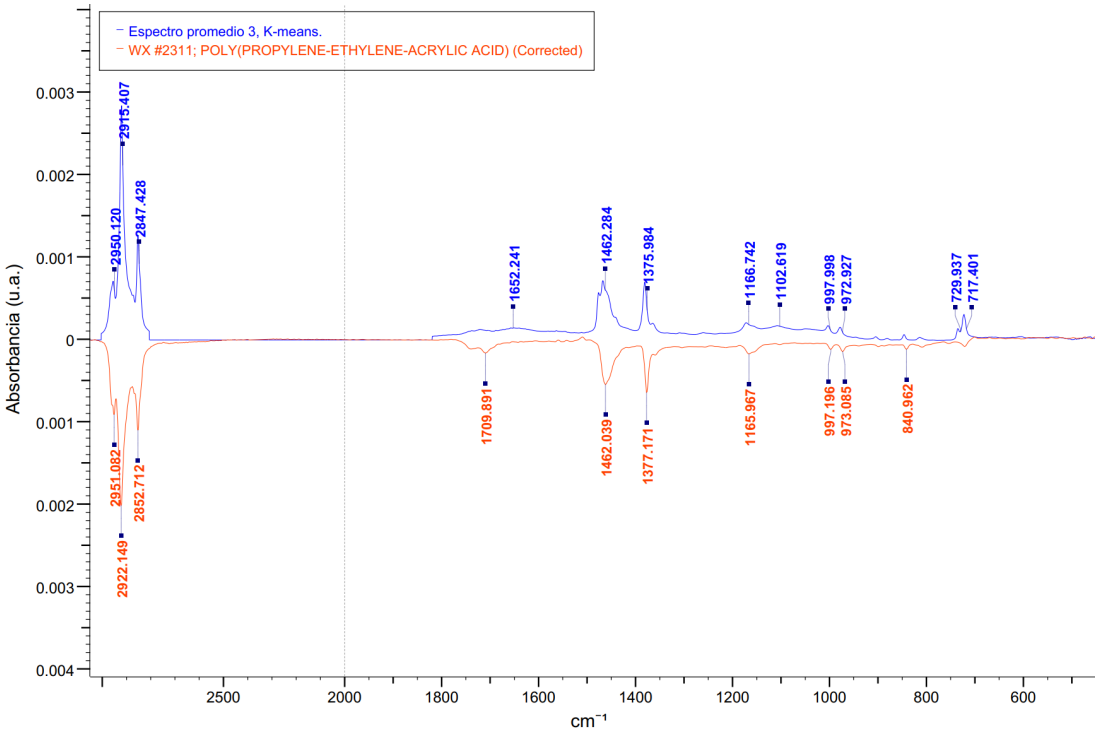
Nota. Comparación de espectro de referencia *HPX 371* vs. espectro promedio 1 (*k-means*) con $HQI = 63{,}23$.

Figura 13(b)



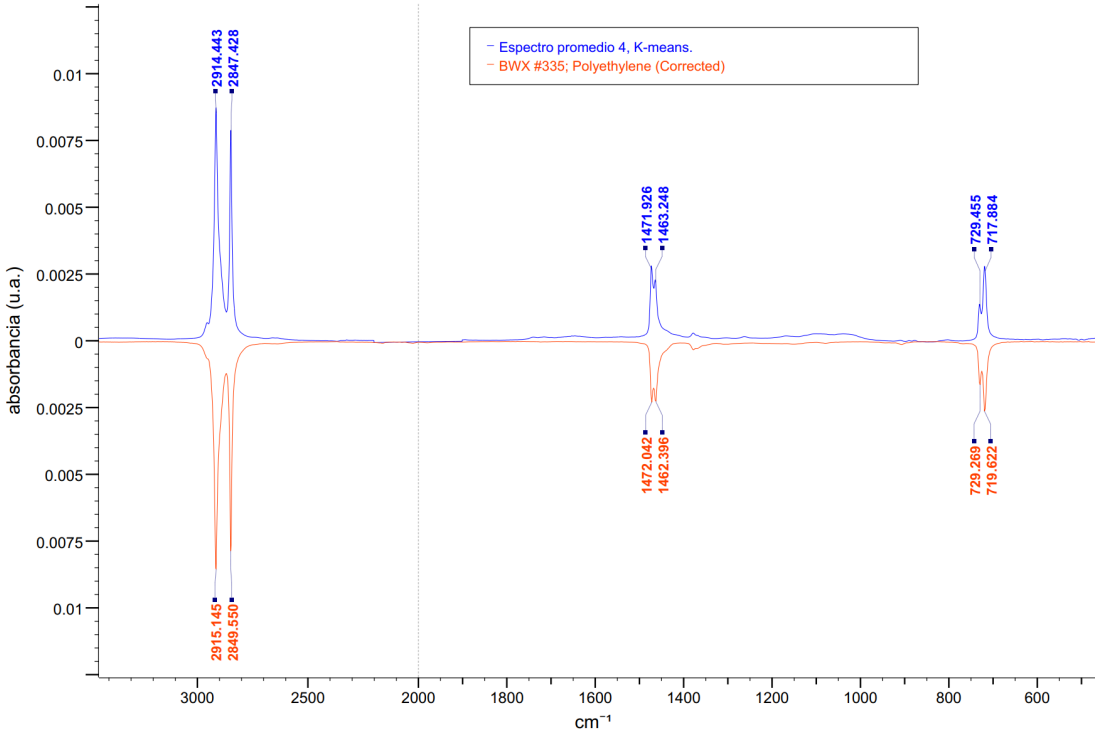
Nota. Comparación de espectro de referencia PET vs. espectro promedio 2 (*k-means*) con HQI=68.37.

Figura 13(c)



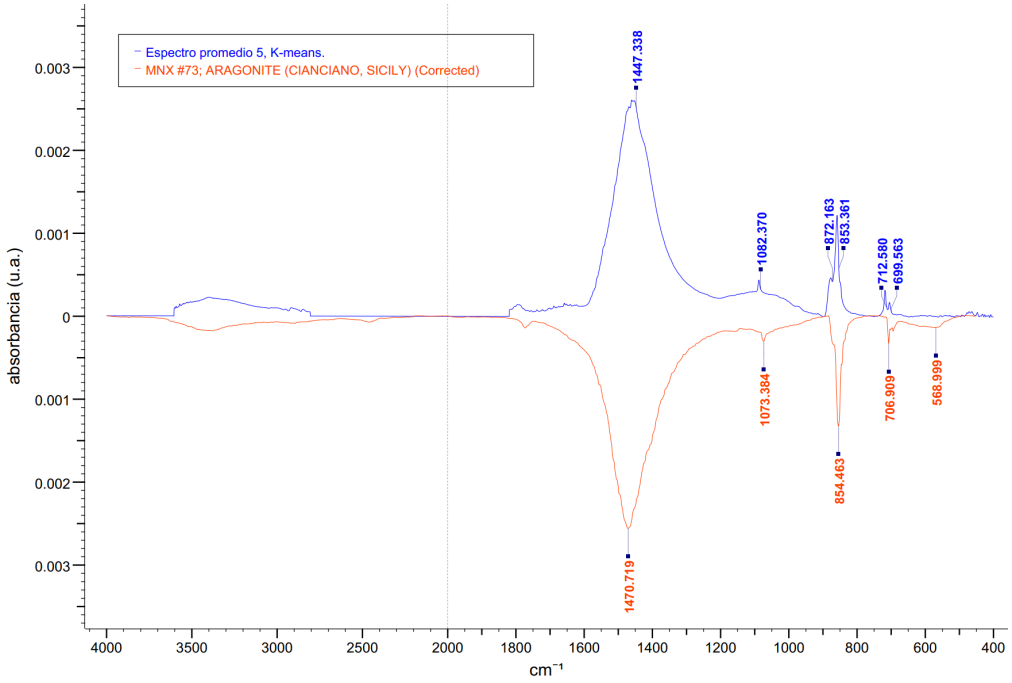
Nota. Comparación de espectro de referencia *WX 2311* vs. espectro promedio 3 (*k-means*) con HQI=81.68.

Figura 13(d)



Nota. Comparación de espectro de referencia Polietileno vs. espectro promedio 4 (*k-means*) con HQI=92.60.

Figura 13(e)



Nota. Comparación de espectro ref. Aragonito vs. espectro promedio 5 (*k-means*) con HQI=86.47

3.5.2 Grupos sugeridos por enfoque jerárquico aglomerativo

Al repetir el ejercicio anterior para los espectros promedio sugeridos por el algoritmo de agrupamiento jerárquico aglomerativo se obtuvo lo consignado en la Tabla 6. Se encontró nuevamente que los primeros dos agrupamientos (A y B) dados por el algoritmo poseen una alta similaridad entre sí y corresponden según la correlación establecida con los espectros de dos diferentes tipos de polietileno.

Tabla 6.
Resultado del ejercicio de adaptación de los espectros promedios obtenidos mediante la estrategia de enfoque jerárquico confrontando la base de datos espectrales KnowItAll.

	Coincidencia	Espectro de referencia Base de datos	HQI
Grupo A	Polietileno	BWX335	92.60
Grupo B	Polietileno de baja densidad	BPX 316	89.20
Grupo C	Poliéster - Algodón	HCX 371	66.15
Grupo D	Aragonito	MNX 73	86.47
Grupo E	Aceite de silicona (PDMS)	WSAAX	94.47

Para el espectro correspondiente al primer agrupamiento la mejor correlación encontrada con la base de datos de referencia corresponde al Polietileno con un HQI de 92.60 (Figura 14(a)) en el cual destacan los ya mencionados picos de estiramiento simétrico – asimétrico y de flexión del grupo metilo, así como la flexión del $C-H$ en grupos metileno. Así mismo, el espectro promedio asociado al segundo agrupamiento se relaciona con polietileno de baja densidad mediante un HQI de 89.20 conservando una misma distribución de picos en su espectro (Figura 14(b)).

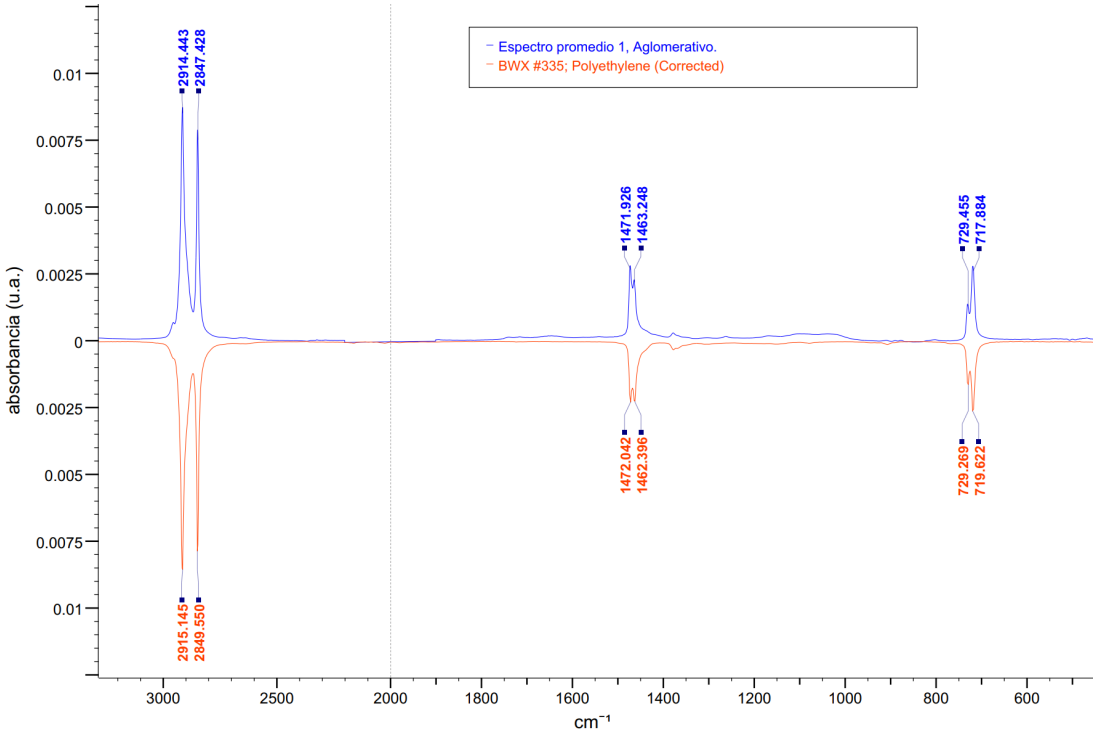
Se observa que el tercer conglomerado (grupo C) sugerido por el algoritmo jerárquico es el más densamente poblado de espectros y agrupa a los que conforman el primer y segundo conglomerado establecidos por la estrategia *k-means* (grupos (a) y (b)), ver Tabla 7 y Tabla 8. El espectro promedio de este agrupamiento (Figura 14(c)) se relaciona con el espectro de referencia de una mezcla en proporción 2:1 de poliéster y fibra textil de algodón con un HQI de 66.15 y se destacan los mismos picos de absorbancia mencionados en el primer conglomerado obtenido por *k-means*.

Es importante resaltar que la alta variabilidad en la intensidad del pico $-OH$ de los espectros individuales que conforman el tercer agrupamiento (grupo C), evidenciado anteriormente en la Figura 14(c), sugiere que algunas de estas muestras, cuyos espectros poseen una intensidad de absorbancia baja en la región de estiramiento del grupo hidroxilo ($3600cm^{-1}$ y $3200cm^{-1}$) se corresponden mejor al PET de tipo resina que a la fibra de poliéster de uso textil; esta separación no se pudo evidenciar mediante el algoritmo jerárquico aglomerativo.

También se encontró que el espectro promedio del cuarto agrupamiento (grupo D) corresponde al mineral Aragonito (Figura 14(d)) con HQI de 86.47. Finalmente, el espectro promedio correspondiente al quinto agrupamiento fue identificado únicamente por el algoritmo jerárquico y se asoció al espectro de referencia del aceite de silicona o poli-dimetil siloxano (PDMS) con un HQI de 94.47, el cual se caracteriza por su banda de vibración de tipo estiramiento del enlace (Si–C) en grupos funcionales metilo en la región $1260cm^{-1}$ y del tipo flexión en la región $1010cm^{-1}$, así como una banda intensa en la región cercana a $780cm^{-1}$ asociada a vibraciones de tipo estiramiento del enlace Si–CH_3 en los grupos metilo.

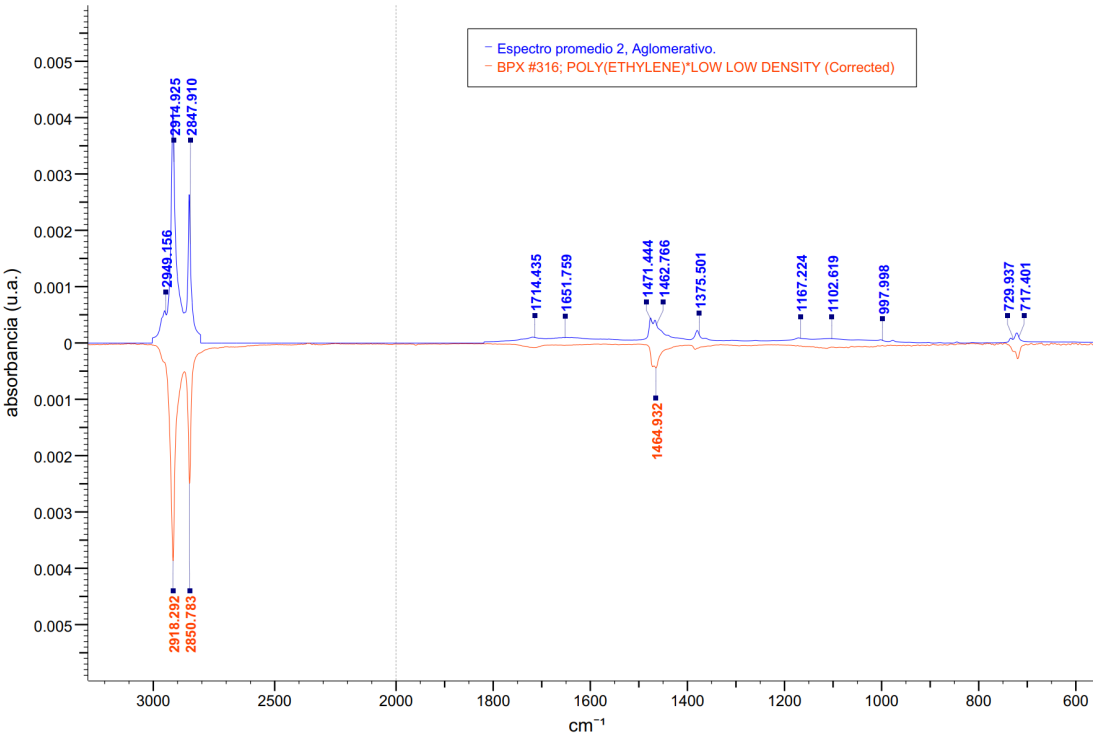
Figura 14.
Adaptación de los espectros promedio sugeridos por el enfoque aglomerativo jerárquico de la base de datos *KnowItAll*.

Figura 14(a)



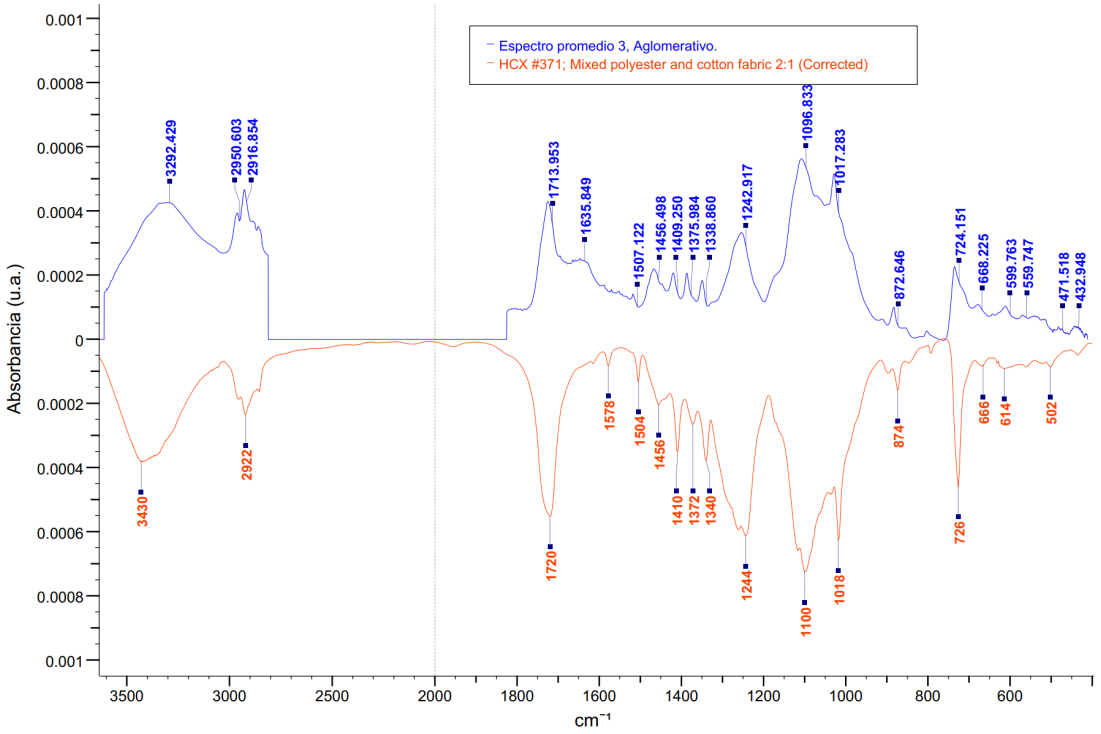
Nota. Comparación de espectro de referencia Polietileno vs. espectro promedio 1 (Aglomerativo) con HQI=92.60.

Figura 14(b)



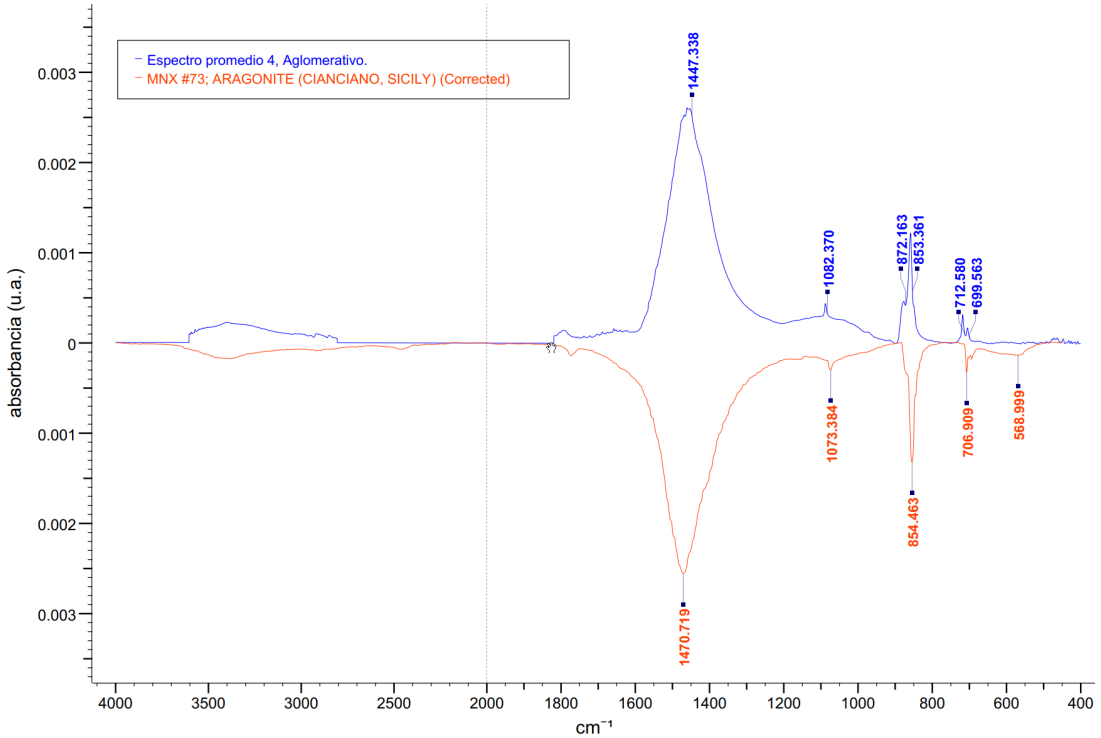
Nota. Comparación de espectro de referencia LLDPE vs. espectro promedio 2 (Aglomerativo) con HQI=89.20.

Figura 14(c)



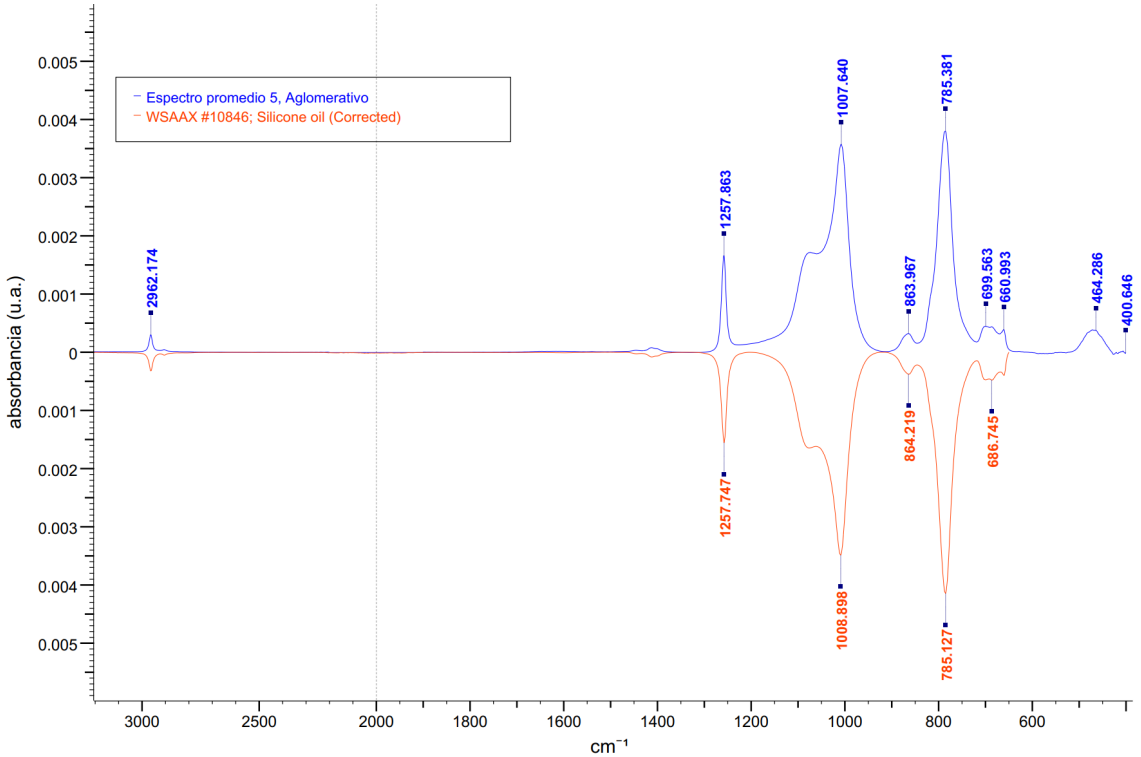
Nota. Comparación de espectro de referencia *HCX 371* vs. espectro promedio 3 (Aglomerativo) con HQI=66.15.

Figura 14(d)



Nota. Comparación de espectro de ref. Aragonito vs. espectro promedio 4 (Aglomerativo) con HQI=86.47.

Figura 14(e)



Nota. Comparación de espectro de referencia Aceite silicona vs. espectro promedio 5 (Aglomerativo) con HQI=94.47.

3.6 Distribución de las muestras en las estaciones de los PNNs

Es de interés en esta monografía establecer la presencia de MPs en las estaciones de monitoreo ubicadas en los Parques Nacionales Naturales Corales de Profundidad y Corales del Rosario y San Bernardo. Para ello se relaciona en la Tabla 7 la distribución geográfica, por estaciones, de los agrupamientos sugeridos mediante la estrategia *k-means*, y en la Tabla 8 se presenta la distribución de los grupos generados por el enfoque jerárquico.

Se observa la presencia de muestras de MPs en todas las estaciones de monitoreo, con una abundancia relativa mayor en la zona costera de Playa Blanca, la cual puede explicarse en función de la alta actividad antropogénica del lugar; sin embargo, se reporta una cantidad significativa de MPs en algunas de las estaciones que se encuentran a más de $50km$ de la línea costera. Una explicación para ello podría estar en las corrientes marinas que arrastran desde la costa el material particulado suspendido en el agua de mar, influenciadas por los vientos que soplan sobre esta región en temporada seca que, al llegar a la cuenca semicerrada del Caribe suroccidental, generan una corriente de circulación ciclónica que gira en sentido contrario de las manecillas del reloj sobre el golfo del Darién, en donde se ubica nuestra área de interés (Andrade, 2001).

Tabla 7. *Distribución por estaciones de las muestras conforme al agrupamiento k-means.*

Estación	Grupo a	Grupo b	Grupo c	Grupo d	Grupo e	Total
1. Playa Blanca	111	26	15	2	3	157
2. Canal Talud	61	26	3	2	7	99
3. Isla Tesoro	49	19	14	1	0	83
4. Bajo Frijol	58	18	4	0	0	80
5. Islote	15	0	3	0	0	18
6. Terraza Sureste	35	6	3	0	0	44
7. Bajo Tortuga	37	20	4	0	0	61
8. Formación CP Sur Talud	67	0	2	0	0	69
9. Costa Sucre	70	2	1	0	1	74
10. Formación CP Centro Talud	45	1	0	0	0	46
11. Isla Mangle	58	7	1	1	0	67
12. Diapiros Noreste	17	1	2	0	0	20
Total	623	126	52	6	11	818

Tabla 8. *Distribución por estaciones de las muestras conforme al enfoque jerárquico aglomerativo.*

Estación	Grupo A	Grupo B	Grupo C	Grupo D	Grupo E	Total
1. Playa Blanca	2	10	141	3	1	157
2. Canal Talud	2	1	89	7	0	99
3. Isla Tesoro	1	11	71	0	0	83
4. Bajo Frijol	0	0	80	0	0	80
5. Islote	0	0	18	0	0	18
6. Terraza Sureste	0	1	43	0	0	44
7. Bajo Tortuga	0	0	61	0	0	61
8. Formación CP Sur Talud	0	0	69	0	0	69
9. Costa Sucre	0	0	73	1	0	74
10. Formación CP Centro Talud	0	0	46	0	0	46
11. Isla Mangle	0	0	66	0	0	67
12. Diapiros Noreste	0	0	20	0	0	20
Total	6	23	777	11	1	818

Respecto al tipo de MP detectado, ambos algoritmos establecen que más del 80 % de las muestras encontradas en cada una de las estaciones corresponden a distintos tipos de Tereftalato de polietileno (como poliéster mezclado con fibras de algodón de uso textil, o como resina de PET para envases y empaques), destacando también la presencia de polietileno (PE) en las estaciones de Playa Blanca e Isla Tesoro.

Resulta de especial importancia la detección de mineral Aragonito, principalmente en la estación canal talud sobre el arrecife coralino, del cual dependen los corales para su crecimiento al ser la materia prima para la formación de su estructura esquelética, mejorando así el hábitat marino y absorbiendo CO_2 atmosférico. También fue detectado por el algoritmo jerárquico aglomerativo en la estación Playa Blanca el polidimetilsiloxano (PDMS) o aceite de silicona, utilizado principalmente como lubricante industrial.

Al comparar la Tabla 7 y Tabla 8 se encuentra que el grupo d se empareja con el grupo A y el grupo e con el grupo D, los cuales se refieren a compuestos identificados como polietileno y aragonito, respectivamente. En ambos casos, se observa que dos algoritmos de agrupamiento diferentes generaron grupos semejantes conformados por exactamente los mismos elementos. Por otra parte, la suma de los elementos que conforman los grupos a y b corresponden de manera aproximada a los del grupo C; esto corresponde al compuesto identificado como poliéster mezclado con fibra de algodón, que es el más común identificado en todas las estaciones de muestreo y representa más del 85 % de las muestras procesadas.

4. Conclusiones

Para muestras de agua de mar recolectadas en doce estaciones del PNN Corales del Rosario y San Bernardo y PNN Corales de Profundidad, se desarrollaron dos modelos de agrupamiento según los algoritmos *k-means* y jerárquico aglomerativo. Los índices de validación interna Dunn, Davies-Bouldin y Silueta simplificado permiten acordar un número de cinco grupos. La confrontación del espectro promedio de cada conjunto sugerido con la base de datos *KnowItAll* (Bio-Rad/Wiley) evidencia la presencia de polímeros, que por su tamaño son catalogados como microplásticos (MPs).

La estrategia *k-means* permite diferenciar dos clases de PET (grupos a y b), en contraste con el enfoque jerárquico aglomerativo que los agrupa en un único conjunto (grupo C). Esto podría explicarse al preagrupamiento que se implementa en *k-means*. En ese sentido, los procedimientos de separación previa impactan positivamente el ejercicio de formación de aglomerados.

Se optimizó el número de grupos para la estrategia *k-means*, pero al cortar el árbol (dendrograma) para generar una cantidad igual de conjuntos se observa la presencia de un grupo (grupo E) con un único elemento, lo cual indica que esta estrategia sugiere una menor cantidad de grupos (posiblemente cuatro o menos).

La distribución de cantidad y tipo de microplásticos encontrados permite asociarlo con la cercanía a comunidades costeras, de relativamente alta dinámica en actividades de turismo y para zonas separadas a más de $50km$ del litoral puede explicarse por la acción de corrientes marítimas.

5. Consideraciones y estudios posteriores

La proyección inmediata es proponer un modelo de clasificación apoyado en el espectro promedio de cada uno de los grupos propuestos en el ejercicio de agrupamiento. De hecho, es la tendencia en este sector, tratar de crear bases de datos de polímeros que están sometidos a las condiciones ambientales del medio marino. Esta exposición genera un proceso de degradación, que se evidencia en las variaciones de intensidad en picos de absorbancia de sus espectros infrarrojos, lo cual dificulta la detección por comparación directa con bases de datos actuales que únicamente aportan referencias de polímeros prístinos. En este sentido, se propone explorar un enfoque similar al *peak search*, centrado en comparar espectros promedio obtenidos, con regiones o picos específicos de la huella vibracional de microplásticos comunes en medio marino, permitiendo automatizar, en una primera aproximación, el proceso de confrontación con base de datos.

Nuestro estudio sugiere la viabilidad de investigaciones futuras en la identificación de microplásticos mediante algoritmos de inteligencia artificial. Se propone explorar técnicas más avanzadas, como el aprendizaje profundo mediante refuerzo, para mejorar aún más la precisión en el agrupamiento y posterior clasificación de microplásticos. Además, la integración de datos multiespectrales y otros tipos de datos, como información sobre corrientes oceánicas y su variación a lo largo del año, podrían proporcionar una comprensión más holística de la distribución y el transporte de microplásticos en los océanos.

La principal limitante encontrada en el desarrollo de este proyecto ha sido la falta de bases de datos amplias y de libre acceso que permitan evaluar en detalle la calidad de los agrupamientos obtenidos, haciendo posible retroalimentar el modelo para lograr mejores resultados.

Se espera concretar cuál es el tipo de MP que amenaza las especies que habitan los Parques Nacionales Naturales Corales de Profundidad y Corales del Rosario y San Bernardo.

Bibliografía

Abidi, N. (2022). *FTIR Microspectroscopy: Selected Emerging Applications*. Springer.

Acosta, I., Duran, M., Rodriguez-Cavallo, E., Mercado-Camargo, J., Mendez-Cuadro, D., and Olivero-Verbel, J. (2019). Quantification of microplastics along the caribbean coastline of colombia: Pollution profile and biological effects on caenorhabditis elegans. *Marine Pollution Bulletin*, 146:574–583.

Acosta, I., Mendez, D., Rodriguez-Cavallo, E., de la Rosa, J., and Olivero-Verbel, J. (2018). Trace elements in microplastics in cartagena: A hotspot for plastic pollution at the caribbean. *Marine Pollution Bulletin*, 139:402–411.

Aggarwal, C. and Reddy, C. K. (2013). *Data Clustering: Algorithms and Applications*. Chapman Hall/CRC Data Mining and Knowledge Discovery Series. Chapman and Hall/CRC, 1st edition.

Alpaydin, E. (2014). *Introduction to Machine Learning*. The MIT Press.

Andrade, C. (2001). Las corrientes superficiales en la cuenca de colombia observadas con boyas de deriva. *Revista de la Academia Colombiana de Ciencias Exactas Físicas y Naturales*, 25:321–335.

Andrady, A. L. (2011). Microplastics in the marine environment. *Marine Pollution Bulletin*, 62(8):1596–1605.

Anger, P. M., von der Esch, E., Baumann, T., Elsner, M., Niessner, R., and Ivleva, N. P. (2018). Raman microspectroscopy as a tool for microplastic particle analysis. *TrAC Trends in Analytical Chemistry*, 109:214–226.

Ansari, Z., Azeem, M. F., Ahmed, W., and Vinaya Babu, A. (2015). Quantitative Evaluation of Performance and Validity Indices for Clustering the Web Navigational Sessions. *arXiv e-prints*, page arXiv:1507.03340.

Banco Mundial (2022). Población, total - colombia.

Bergmann, M., Gutow, L., and Klages, M. e. (2015). *Marine Anthropogenic Litter*. Springer International Publishing, 1 edition.

Blum, M.-M. and John, H. (2012). Historical perspective and modern applications of attenuated total reflectance – fourier transform infrared spectroscopy (atr-ftir). *Drug Testing and Analysis*, 4(3-4):298–302.

Bonaccorso, G. (2018). *Machine Learning Algorithms Popular algorithms for data science and machine learning*. Packt.

Celebi, M. E. e. (2015). *Partitional Clustering Algorithms*. Springer International Publishing, 1 edition.

Chen, X., ming Yuan, L., Yi, G., Huang, G., Shi, W., and Chen, X. (2022). A rapid automatic spectroscopic identification method of environmental microplastics. *Chemometrics and Intelligent Laboratory Systems*, 222:104511.

Davies, D. and Bouldin, D. (1979). A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, PAMI-1:224 – 227.

Desforges, J.-P. W., Galbraith, M., and Ross, P. S. (2015). Ingestion of microplastics by zooplankton in the northeast pacific ocean. *Archives of Environmental Contamination and Toxicology*, 69(3):320–330.

Dong, M., She, Z., Xiong, X., Ouyang, G., and Luo, Z. (2022). Automated analysis of microplastics based on vibrational spectroscopy: are we measuring the same metrics? *Analytical and Bioanalytical Chemistry*, 414(11):3359–3372.

Dougherty, G. (2012). *Pattern Recognition and Classification: An Introduction*. Springer, 2013 edition.

Dunn, J. C. (1973). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3(3):32–57.

Forsyth, D. (2019). *Applied Machine Learning*.

Gago, J., Carretero, O., Filgueiras, A., and Viñas, L. (2018). Synthetic microfibers in the marine environment: A review on their occurrence in seawater and sediments. *Marine Pollution Bulletin*, 127:365–376.

Gago, J., Galgani, F., Maes, T., and Thompson, R. C. (2016). Microplastics in seawater: Recommendations from the marine strategy framework directive implementation process. *Frontiers in Marine Science*, 3.

Garcés-Ordóñez, O., Espinosa, L. F., Costa Muniz, M., Salles Pereira, L. B., and Meigikos dos Anjos, R. (2021). Abundance, distribution, and characteristics of microplastics in coastal surface waters of the colombian caribbean and pacific. *Environmental Science and Pollution Research*, 28(32):43431–43442.

Geyer, R., Jambeck, J. R., and Law, K. L. (2017). Production, use, and fate of all plastics ever made. *Science Advances*, 3(7):e1700782.

Halkidi, M., Batistakis, Y., and Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, 17(2):107–145.

Han, J., Kamber, M., and Pei, J. (2011). *Data Mining. Concepts and Techniques, 3rd Edition*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 3rd edition edition.

Jambeck, J. R., Geyer, R., Wilcox, C., Siegler, T. R., Perryman, M., Andrady, A., Narayan, R., and Law, K. L. (2015). Plastic waste inputs from land into the ocean. *Science*, 347(6223):768–771.

Jiang, Y., Yang, F., Hassan Kazmi, S. S. U., Zhao, Y., Chen, M., and Wang, J. (2022). A review of microplastic pollution in seawater, sediments and organisms of the chinese coastal and marginal seas. *Chemosphere*, 286:131677.

Johnston, B., Jones, A., and Kruger, C. (2019). *Applied Unsupervised Learning with Python: Discover hidden patterns and relationships in unstructured data with Python*. Packt Publishing, 1st edition edition.

Jung, M. R., Horgen, F. D., Orski, S. V., Rodriguez C., V., Beers, K. L., Balazs, G. H., Jones, T. T., Work, T. M., Brignac, K. C., Royer, S.-J., Hyrenbach, K. D., Jensen, B. A., and Lynch, J. M. (2018). Validation of atr ft-ir to identify polymers of plastic marine debris, including those ingested by marine organisms. *Marine Pollution Bulletin*, 127:704–716.

Käppler, A., Fischer, D., Oberbeckmann, S., Schernewski, G., Labrenz, M., Eichhorn, K.-J., and Voit, B. (2016). Analysis of environmental microplastics by vibrational microspectroscopy: Ftir, raman or both? *Analytical and Bioanalytical Chemistry*, 408(29):8377–8391.

Kaufman, L. and Rousseeuw, P. J. (2005). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, 1 edition.

Kaur, H. (2021). Fundamentals of atr-ftir spectroscopy and its role for probing in-situ molecular-level interactions. In D. K. Singh, M. Pradhan, . A. M., editor, *Modern Techniques of Spectroscopy: Basics, Instrumentation, and Applications.*, pages 1804–1819. Springer Nature Singapore.

Kovač, M., Palatinus, A., Koren Bačovnik, , Peterlin, M., Horvat, P., and Kržan, A. (2016). Protocol for microplastics sampling on the sea surface and sample analysis. *Journal of Visualized Experiments*, 2016.

Lenssen, L. and Schubert, E. (2022). Clustering by direct optimization of the medoid silhouette. In Skopal, T., Falchi, F., Lokoč, J., Sapino, M. L., Bartolini, I., and Patella, M., editors, *Similarity Search and Applications*, pages 190–204, Cham. Springer International Publishing.

Liu, Y., Li, Z., Xiong, H., Gao, X., and Wu, J. (2010). Understanding of internal clustering validation measures. In *Proceedings of the 2010 IEEE International Conference on Data Mining*, ICDM '10, page 911–916, USA. IEEE Computer Society.

Maulik, U. and Bandyopadhyay, S. (2002). Performance evaluation of some clustering algorithms and validity indices. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1650–1654.

Milosevic, M., M. F. V. (2012). *Internal Reflection and ATR Spectroscopy*, volume Vol. 176 of *Chemical Analysis: A Series of Monographs on Analytical Chemistry and Its Applications*.

Miri, S., Saini, R., Davoodi, S. M., Pulicharla, R., Brar, S. K., and Magdouli, S. (2022). Biodegradation of microplastics: Better late than never. *Chemosphere*, 286:131670.

Mukherjee, S., Martínez-González, J. , Dowling, D. P., and Gowen, A. A. (2018). Predictive modelling of the water contact angle of surfaces using attenuated total reflection – fourier transform infrared (atr-ftir) chemical imaging and partial least squares regression (plsr). *Analyst*, 143:3729–3740.

Nakano, H. and Arakawa, H. (2022). Oceanic microplastics in japan: A brief review on research protocol and present pollution. *Regional Studies in Marine Science*, 51:102201.

Ministerio de Ambiente y desarrollo sostenible (2019). Asuntos marinos, costeros y recursos acuáticos.

Ministerio de Ambiente y desarrollo sostenible (2021). Arrecifes de coral, un patrimonio que colombia restaura y conserva.

Oberbeckmann, S., Loeder, M. G., Gerdts, G., and Osborn, A. M. (2014). Spatial and seasonal variation in diversity and structure of microbial biofilms on marine plastics in Northern European waters. *FEMS Microbiology Ecology*, 90(2):478–492.

Otto, M. (2016). *Pattern Recognition and Classification*, chapter 5, pages 135–211. John Wiley Sons, Ltd.

Primpke, S., Wirth, M., Lorenz, C., and Gerdts, G. (2018). Reference database design for the automated analysis of microplastic samples based on fourier transform infrared (ftir) spectroscopy. *Analytical and Bioanalytical Chemistry*, 410(21):5131–5141.

Reddy, C. K. and Vinzamuri, B. (2021). A survey of partitional and hierarchical clustering algorithms. In Reddy, C. K. and Aggarwal, C. C., editors, *Data Clustering: Algorithms and Applications.*, pages 1804–1819. CRC Press.

Renner, G., Nellessen, A., Schwiers, A., Wenzel, M., Schmidt, T. C., and Schram, J. (2019). Data preprocessing evaluation used in the microplastics identification process: A critical review practical guide. *TrAC Trends in Analytical Chemistry*, 111:229–238.

Renner, G., Schmidt, T. C., and Schram, J. (2017). A new chemometric approach for automatic identification of microplastics from environmental compartments based on ft-ir spectroscopy. *Analytical Chemistry*, 89(22):12045–12053.

Robin, R., Karthik, R., Purvaja, R., Ganguly, D., Anandavelu, I., Mugilarasan, M., and Ramesh, R. (2020). Holistic assessment of microplastics in various coastal environmental matrices, southwest coast of india. *Science of The Total Environment*, 703:134947.

Rocha-Santos, T., Costa, M. F., and Mouneyrac, C. (2022). *Handbook of Microplastics in the Environment*. Springer Nature Reference. Springer.

Seghers, J., Stefaniak, E. A., La Spina, R., Cella, C., Mehn, D., Gilliland, D., Held, A., Jacobsson, U., and Emteborg, H. (2022). Preparation of a reference material for microplastics in water—evaluation of homogeneity. *Analytical and Bioanalytical Chemistry*, 414(1):385–397.

Shim, W. J., Hong, S. H., and Eo, S. E. (2017). Identification methods in microplastic analysis: a review. *Anal. Methods*, 9:1384–1391.

Skoog, D. A. (2019). *Principios de análisis instrumental*. Cengage Learning, 7 edition.

Socrates, G. (2001). *Infrared and Raman characteristic group frequencies: tables and charts*. Wiley, 3 edition.

Tharwat, A., Gaber, T., Ibrahim, A., and Hassanien, A. E. (2017). Linear discriminant analysis: A detailed tutorial. *AI Communications*, 30:169–190. 2.

Tosin, M., Weber, M., Siotto, M., Lott, C., and Degli-Innocenti, F. (2012). Laboratory test methods to determine the degradation of plastics in marine environmental conditions. *Frontiers in Microbiology*, 3.

Walsh, A. N., Reddy, C. M., Niles, S. F., McKenna, A. M., Hansel, C. M., and Ward, C. P. (2021). Plastic formulation is an emerging control of its photochemical fate in the ocean. *Environmental Science & Technology*, 55(18):12383–12392.

Weisser, J., Pohl, T., Heinzinger, M., Ivleva, N. P., Hofmann, T., and Glas, K. (2022). The identification of microplastics based on vibrational spectroscopy data – a critical review of data analysis routines. *TrAC Trends in Analytical Chemistry*, 148:116535.

Wesch, C., Bredimus, K., Paulus, M., and Klein, R. (2016). Towards the suitable monitoring of ingestion of microplastics by marine biota: A review. *Environmental Pollution*, 218:1200–1208.

Zhang, K., Hamidian, A. H., Tubić, A., Zhang, Y., Fang, J. K., Wu, C., and Lam, P. K. (2021). Understanding plastic degradation and microplastic formation in the environment: A review. *Environmental Pollution*, 274:116554.

Zhang, Y., Wu, H., Xu, L., Liu, H., and An, L. (2022). Promising indicators for monitoring microplastic pollution. *Marine Pollution Bulletin*, 182:113952.

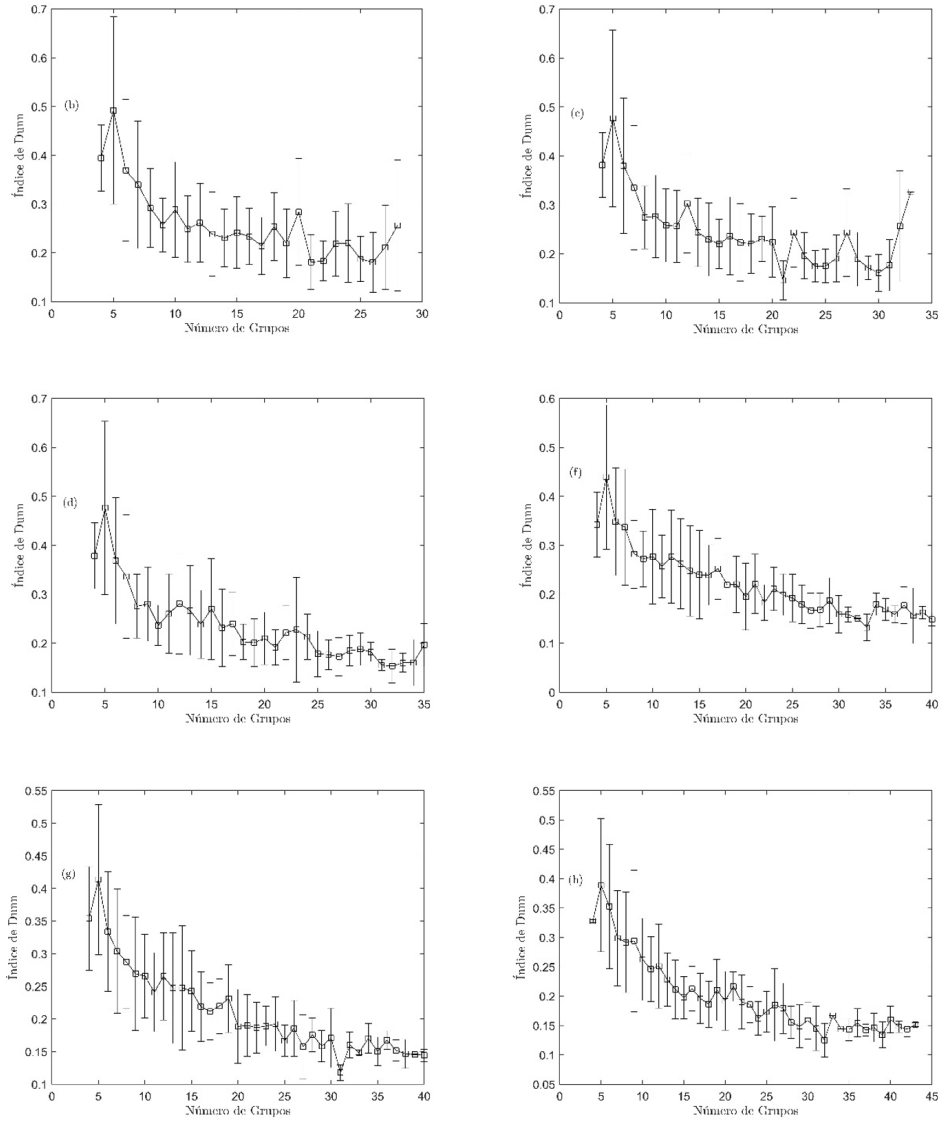
Zhou, Z.-H. (2021). *Machine Learning*. Springer Singapore.

Zhu, S., Chen, H., Wang, M., Guo, X., Lei, Y., and Jin, G. (2019). Plastic solid waste identification system based on near infrared spectroscopy in combination with support vector machine. *Advanced Industrial and Engineering Polymer Research*, 2(2):77–81.

A. Anexos

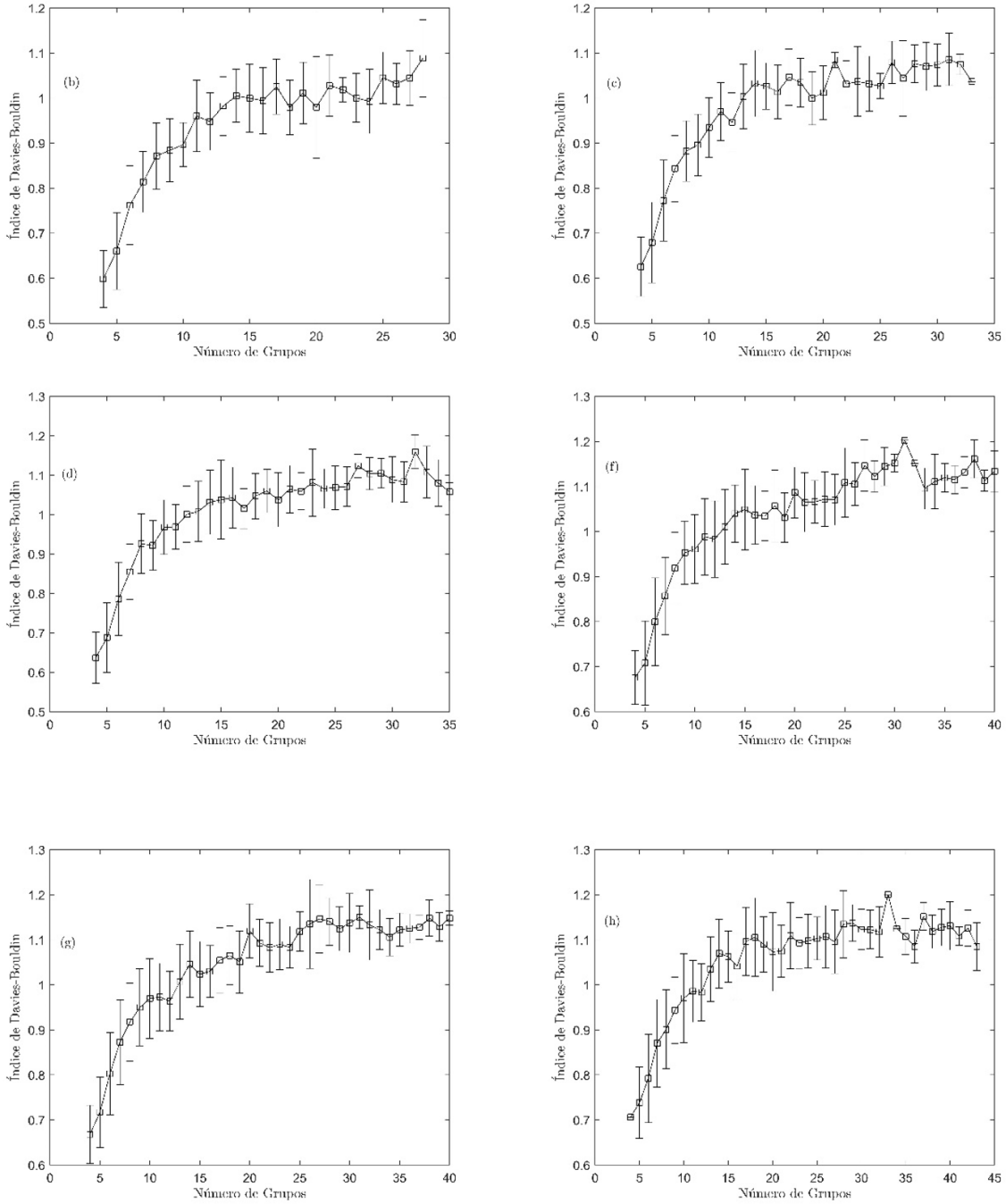
A.1 Anexo A. Índice de Dunn

Figura 15.
Curva de índices de Dunn versus número de grupos. En (b), (c), (d), (f), (g) y (h), se consideran ocho (8), nueve (9), diez (10), doce (12), trece (13) y catorce (14) componentes principales, respectivamente.



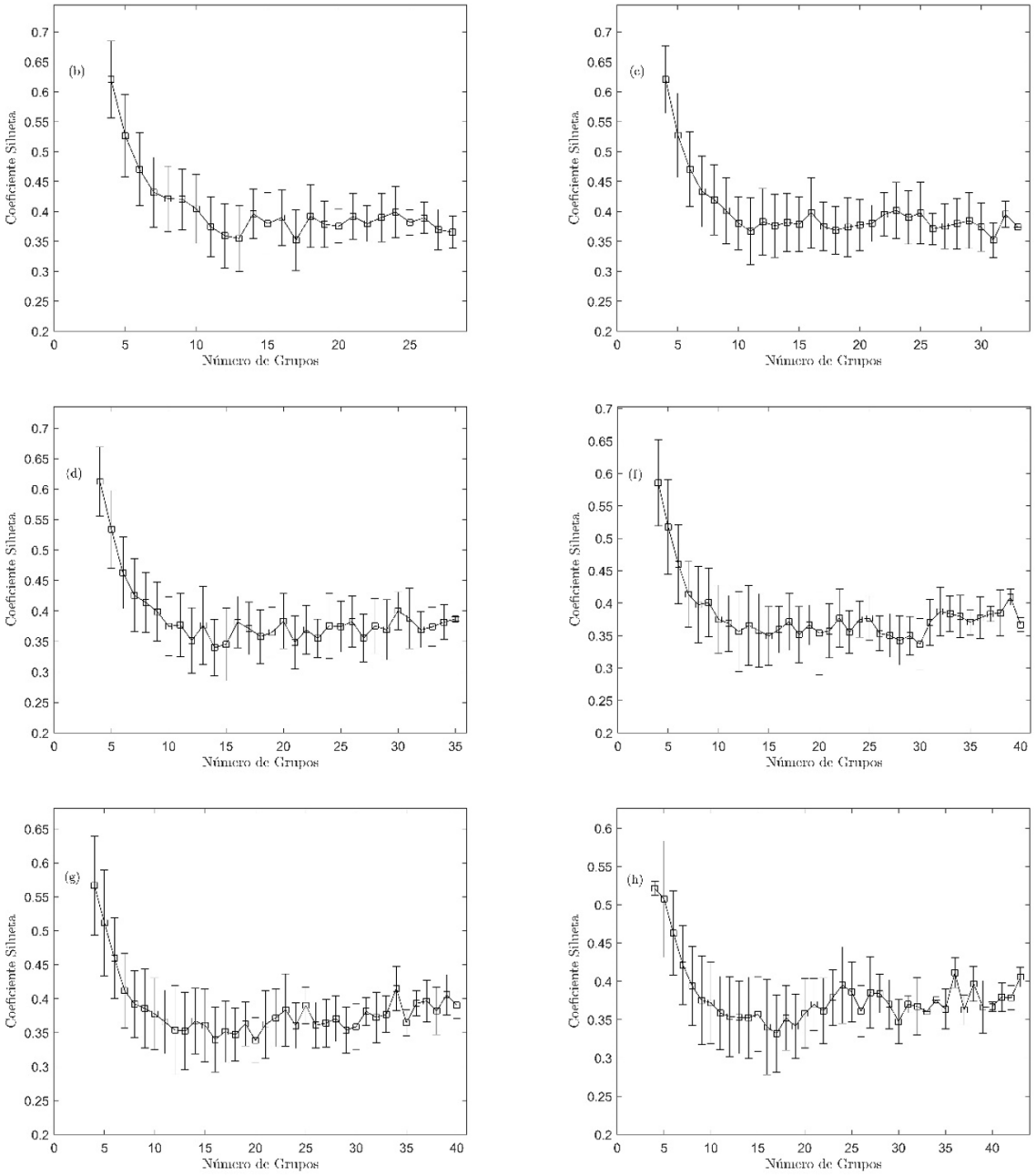
A.2 Anexo B. Índice de Davies - Bouldin

Figura 16.
Curva de índices de Davies - Bouldin versus número de grupos. En (b), (c), (d), (f), (g) y (h), se consideran ocho (8), nueve (9), diez (10), doce (12), trece (13) y catorce (14) componentes principales, respectivamente.



A.3 Anexo C. Coeficiente Silueta simplificado

Figura 17.
Curva de coeficiente de siluetas versus número de grupos. En (b), (c), (d), (f), (g) y (h), se consideran ocho (8), nueve (9), diez (10), doce (12), trece (13) y catorce (14) componentes principales, respectivamente.



A.4 Anexo D. Códigos implementados en Matlab R2019b

Programa principal

```
    clc     %%% Borrar pantalla
    clear %%% Borrar variables previas en el espacio de trabajo
    close all %%% Cerrar ventanas

for a = 7 %%% Numero de PCs
    cd( C :\Users\Escritorio\Rutinas_MP_Abril )
    tic     %%% Se mide tiempo de ejecucion.
    %%% Llamado de los archivos
    Xtodos = [];   Stodos = [];    Esttodos = [];
    NEstaciones = 12;    %%%Numero de estaciones
    sf = 1;  %%% 2 para visualizar graficas, 1 para evitar su despliegue
    for Estacion = 1:NEstaciones
        z=[ C :\Users\Escritorio\Rutinas_MP_Abril\CORRECCION LINEA BASE - ESTACION
             num2str(Estacion)];
        Carpeta = z;
        archivos1 = fileDatastore(Carpeta,    ReadFcn    ,@customreader,
             IncludeSubfolders    ,1,    FileExtensions    ,    . s p a ) ;
        S = size(archivos1.Files);
        for i = 1:S(1)
            fid = fopen(archivos1.Files{i},    r    );
            [s, Wn,~,~] = LoadSpectra(fid);  %%% cargando el archivo
            zr = zeros(size(s));
            s(Wn<2200 & Wn>1900) = zr(Wn<2200 & Wn>1900);  %%% Region anomala del
                diamante
            dl = abs(mean(diff(Wn)));    %%% normalizacion por el area bajo la curva
                espectral
            s = abs(s./areanueva(s,dl));    %%% normalizacion.
            x(i,:) = s;
            Est(i) = Estacion;
        end
        xtodos = [xtodos; x];
        Esttodos = [Esttodos; E s t ];
        Stodos = [Stodos; archivos1.Files];
        clear x Est
    end
    % Sxtodos = size(xtodos);

    %%%%%Determinacion Razon Se al a Ruido
% % % % % %     for Espe = 1:800
% % % % % %     Senal = xtodos(Espe,4400:4500);
% % % % % %     Sprom = mean(Senal);
% % % % % %     desvsenal = std(Senal);
% % % % % %     SenalRuido(Espe) = Sprom/desvsenal;
% % % % % %     figure(1)
% % % % % %     h = gca;
% % % % % %
% % % % % %     plot(Wn,xtodos(Espe,:),Wn(4400:4500),Senal,    r    )
% % % % % %     h.XDir =     reverse    ;
% % % % % %     pause(1)
% % % % % %     end
% % % % % %      mean(SenalRuido)
% % % % % %      std(SenalRuido)
% % % % % %     return

    %%%%%% Reduccion de dimensionalidad en el sentido de PCA
    [V,T,pcvar ] = nipals(xtodos,a);  %%%Algoritmo Nipals.
    % % [T,score,latent,tsquared,explained,mu] = pca(xtodos,    NumComponents    ,2,
            Algorithm    ,    svd    );
```

```
54      %%% V son los valores propios que resultan de la descomposicion en
55      %%% componentes principales
56      %%% T, los vectores propios
57
58      vs = 85;    %%79:95;%%% Coeficiente de similaridad de acuerdo al producto punto
            85
59      for cc = 1:length(vs)
60          rng(89);  %%% Generador de n meros aleatorios
61          Numero = 50;  %%Se ensaya un determinado N mero de modelos
62          randomico = randi(3456,1,Numero);
63          for modelo = 47  %%1:length(randomico)  %%47 para K=5 %%%% %%%40 para 7
64              E = xtodos*pinv(T(:,1:a)   );  %%%Espectros en el espacio creado por
                    PCA
65              R = E;
66              sE = size(E);
67              k = 1;
68              cont = 1;
69              NE = [];
70              rng(randomico(modelo))   %%%Semilla aleatoria, garantiza
                    reproducibilidad.
71              u = 1:sE(1); %% baraja inicio para seleccionar diferentes cabezas de
                    grupo
72              u = circshift(u,-randi(sE(1)));
73              s = 1;  %% contador
74              for i = 1:sE(1)
75                  m = 1;
76                  for j = i+1:sE(1)
77                      if E(u(j),1) ~= -1000
78                          cs = dot(E(u(i),:),E(u(j),:))/(norm(E(u(i),:))*norm(E(u(j)
                                ,:)));  %%% coseno como m trica de similaridad
79                          if (cs> = (vs(cc)/100) &&  cs<=1)
80                              Muestra(s) = u(j);
81                              s = s+1;
82                              m = m+1;
83                              E(u(j),1) = -1000;
84                          end
85                      end
86                  end
87
88                  if m ~= 1
89                      NE(cont,:) = [u(i) m-1];  %%%Cabeza de grupo y n mero de
                            muestras en el grupo.
90                      %%%Las muestras se encuentran en el arreglo Muestra(s).
91                      cont = cont+1;
92                  end
93              end
94
95              EA = find(E(:,1) ~= -1000); %%%Explorando elementos que quedan por
                    fuera
96              Xm1 = R(~ismember(EA,NE(:,1)),:);
97              sX1 = size(Xm1);
98              aa = [0; cumsum(NE(:,2))];  %%% indica n mero acumulado de muestras.
99              sNE = size(NE);
100             grupo = [];
101             for contaMues = 1:sNE(1)
102                 Mues = [NE(contaMues,1); Muestra(aa(contaMues)+1:aa(contaMues+1))
                        ];
103                 f{contaMues} = Mues;
104                 A = [contaMues*ones(length(Mues),1) R(Mues,:)];
105                 M(contaMues,:) = mean(A);
106                 grupo = [grupo; A];
107             end
108             Xm = [ones(sX1(1),1)*(sNE(1)+1), Xm1];
```

```
            grupo = [grupo; Xm];
            MM = M(:,2:end);
            nga = sNE(1);
            if nga > 1
                [Rng,RD,e] = Microplasticos_kmeans(nga,MM,R,sf,Esttodos,Wn,Indi,T,a
                    ,xtodos,NE);
                ER(modelo,:) = [modelo,Rng,RD]; %%% Rng (n mero de grupos) RD,
                     ndice  de Dunn
            else
            end
        end
    end
    euc = sum(e,2);
    e = [e, euc];
    e = [e; sum(e)];
    toc
end

function customreader(filename)
end

function A=areanueva(s,dl)
A=0;
for i = 1:length(s)-1; A = A+dl*(s(i)+s(i+1))/2;end
end

}
```

Para organización y despliegue de resultados

```
    function [nga, Indice, e] = Microsplaticos_kmeans(nga,MM,E,sf,Esttodos,Wn,Indi,
        T,a,xtodos,NE)

NEstaciones = 12;   %%N mero de Estaciones
DatosI = MM;

for cont = 1:length(nga)
    ng = nga(cont);
    [grupos,M] = kmeansj(E,ng,DatosI);

    switch Indi
        case 1
            Indice = Dun(grupos,M);  %%%Dunn Index
        case 2
            Indice = DBI(grupos,M); %%%Davies-Bouldin Index
        case 3
            Indice = siluetaS(grupos,M); %%%Silueta
    end
    Ne = [];
    Indxa = [];
    Grupo = [];
    for k = 1:ng
        indx = find(grupos==k);
        Ne = [Ne, length(indx)];
        Indxa = [indxa; indx , grupos(indx,1)];
        Epg = Esttodos(indx);
        for i = 1:NEstaciones
            Ep =(Epg==i);
            e(i,k) = sum(Ep);
        end
        grupo = [grupo; grupos(indx,:)];
```

```
31
32          if sf == 2
33              meanxtodos = mean(E(indx,:));
34              % %                 meanxtodos =(E(indx,:));
35              %                 Medios(k,:) = (T(:,1:a)*meanxtodos )    ;
36              Medios(k,:) = mean(xtodos(indx,:));
37              MM = [Wn ; Medios(k,:)];
38              nf = [    Medios_7_     , num2str(k)     .xlsx ];
39              xlswrite(nf, MM  );
40              drawnow
41              figure;
42              h = gca;
43
44              plot( Wn  ,xtodos(indx,:)    ,    color    ,[0.9 0.9 0.9]);hold on
45              % % % %                 h1 = plot( Wn  ,max(xtodos(indx,:)),    r    ,
                    LineWidth    ,0.8);
46              h2 = plot( Wn  ,Medios(k,:)    ,    k    ,    LineWidth    ,1);
47              % % %                 h3=plot( Wn  ,min(xtodos(indx,:)),    b    ,
                    LineWidth    ,0.8);
48              h4 = plot( Wn  ,xtodos(NE(k),:)    ,    r    ,    LineWidth    ,1);
49              %                 legend([h1; h2; h3],{    Mximo     ;    Medio    ;
                     Mnimo     },    Box    ,    off    ,    Location    ,
                     northoutside     ,    Orientation    ,    horizontal    );
50              legend([h2; h4],{    Medio    ;  Espectro    Semilla },    Box    ,
                    off    ,    Location    ,    northoutside    ,    Orientation    ,
                     horizontal    );
51              h.XDir =     reverse    ;
52              ylabel(  Absorbancia    (u.a.)    ,    interpreter    ,    latex    );
53              xlabel(  $$   1/\lambda    (  cm ^{-1}) $ $  ,    interpreter    ,
                    latex    )
54              hold off
55              zmax = max(max(xtodos(indx,:)));
56              axis([500 4000 0 max(max(xtodos(indx,:)))]);
57              zz = [    espectro     num2str(k)];
58              etiquetas = char(    (a)    ,    (b)    ,    (c)    ,    (d)    ,    (e)    ,    (
                    f)    ,    (g)    ,    (h)    ,    (i)    ,    (j)    );
59              text(3700,0.9*zmax,etiquetas(k,:))
60              print(num2str(k),    -djpeg ,    -r600 );
61
62          else
63          end
64
65      end
66          silueta_simplificado_grafica(grupo,M)
67          %%%print(    silueta    ,    -djpeg ,    -r1200 );
68  end
```

Índice de Dunn

```
1       function Dunn = Dun(grupo,M)
2   SM = size(M);
3   a = 1:SM(1);
4   M = [ a   ,M];
5   Sg = size(grupo);
6   A = [];
7   for j = 1:SM(1)
8       cc = 1;
9       for i = 1:Sg(1)
10          if grupo(i,1) == M(j,1)
11              A(cc) = distancia(grupo(i,2:end),M(j,2:end));
12              cc = cc+1;
```

```
        else
        end
    end
    Diam(j) = max(A);  %%%Qu  tan Compacto
end
dd = 1;
for i = 1:SM(1)
    for j = i+1:SM(1)
        B(dd) = distancia(M(i,2:end),M(j,2:end));
        dd = dd+1;
    end
end
 Dunn = min(B)/max(Diam);
end
```

Índice de Davies-Bouldin

```
    function DBIn = DBI(grupo,M)
SM = size(M);
a = 1:SM(1);
M = [ a  ,M];
Sg = size(grupo);
for j = 1:SM(1)
    cc = 1;  d=0;
    for i = 1:Sg(1)
        if grupo(i,1) == M(j,1)
           d = d+distancia(grupo(i,2:end),M(j,2:end));
           cc = cc+1;
        else
        end
    end
    A(j) = d./(cc-1);
end
for i = 1:SM(1)
    for j = i+1:SM(1)
        B(i,j) = distancia(M(i,2:end),M(j,2:end));
    end
end

for i = 1:SM(1)
    for j = i+1:SM(1)
        R(i,j) = (A(i)+A(j))./B(i,j);
    end
end
DBIn = mean(max );
end
```

Coeficiente Silueta Simplificado

```
    function CS = Siluetas(grupo,M)
SM = size(M);
EM = 1:SM(1);
M = [ E M  ,M];
Sg = size(grupo);

for i = 1:Sg(1)
    dd = 1;
    for j = 1:SM(1)
```

```
        if grupo(i,1) == M(j,1)
            A = distancia(grupo(i,2:end),M(j,2:end));
        else
            BB(dd) = distancia(grupo(i,2:end),M(j,2:end));
            dd = dd+1;
        end
    end
    b = min(BB);
    a = A;
    s(i,: ) = [grupo(i,1) (b-a)/max(a,b)] ; %% Con producto punto colocar a-b, en caso de distancia euclideana b-a.
end
CA1 = [];
ng = SM(1);
for n = 1:ng
    A = sort(s(s(:,1)==n,2));
    H(n) = mean(A);
    CA = [A; zeros(10,1)];
    L(n) = length(CA);
    CA1 = [CA1;CA];
    clear CA
end
CS = max(mean(H));
end
```

Perfil silueta

```
    function silueta_simplificado_grafica(grupo,M,Indice)
SM = size(M);EM = 1:SM(1);M = [ E M  ,M];Sg = size(grupo);

for i = 1:Sg(1)
    dd = 1;
    for j = 1:SM(1)
        if grupo(i,1) == M(j,1)
            A = distancia(grupo(i,2:end),M(j,2:end));
            % %             A = dot(grupo(i,2:end),M(j,2:end));
        else
            BB(dd) = distancia(grupo(i,2:end),M(j,2:end));
            % % %             BB(dd) = dot(grupo(i,2:end),M(j,2:end));
            dd = dd+1;
        end
    end
    b = min(BB);
    a = A;
    s(i,:) = [grupo(i,1) (b-a)/max(a,b)];  %% Con producto punto colocar a-b, en caso de distancia euclideana b-a.
end

color = [1 0 0;  0 1 0; 0 0 1; 1 1 0; 0 1 1; 1 0 1; 0.9 0.9 0; 0.8 0.8 0.8; 0.5 0 0];
ng = SM(1);
CA1 = [];
for n = 1:ng
    A = sort(s(s(:,1)==n,2));
    H(n) = mean(A);
    CA = [A; zeros(10,1)];
    L(n) = length(CA);
    CA1 = [CA1;CA];
    clear CA
end

```

```
CS = max(mean(H));
L = [1 cumsum(L)];

figure
h1 = gca;
b = barh(CA1);hold on
```

k-means

```
    function [grupos,M,iter]=kmeansj(E,ng,M)
% % % M = Minicial(E,ng);
S = size(E);
  psilon   = 1e-3; %%%%Tolerancia
nr = randi(S(1),[1,S(1)]);  %%% Asignar n meros de etiquetas a las muestras.
a1 = round((S(1)/ng)-0.5);
Mold = M;
iter = 1;
A = ones(1,ng);
while any(A)
    for kk = 1:ng
        for i = 1:S(1)
            dm(kk,i) = distancia(E(i,:),M(kk,:));
        end
    end
    MM = M; M = [];
    [u,v] = find(dm==min(dm));
    for i = 1:ng
        gg = E(v(u(1) == i),:);
        Sgg = size(gg);
        if Sgg(1) > 1
            mgg = mean(gg,    omitnan    );
        else
            mgg = MM(i,:);
        end
        M = [M; mgg];
    end
    for i = 1:ng
        if distancia(Mold(i,:),M(i,:)) < epsilon
            A(i) = 0;
        else
        end
    end
    Mold = M;
    iter = iter+1;
end

for i = 1:S(1)
    for kk = 1:ng
        dist(kk) = distancia(E(i,:),M(kk,:));
    end
    if isnan(dist)
    else
        mindist = min(dist);
        md = find(dist==mindist);
        grupos(i,:) = [md(1) E(i,:)];  %%%Agrupa
    end
end
end
```

NIPALS

```
    function [ T,P,pcvar ] = nipals(X,a,it,tol )
%Nipals algorithm for Principle Component Analysis
%
% Author: Qiaonan Duan, 6/7/2013, MSSM.
%
%
if nargin == 2
    it = 1000;
    tol = 1e-4;
elseif nargin == 3
    tol = 1e-4;
end

[obsCount,varCount] = size(X);
Xh = X;
T = zeros(obsCount,a);
P = zeros(varCount,a);
pcvar = [];

varTotal = sum(var(Xh));
currVar = varTotal;
nr = 0;

for h = 0:a
    th = Xh(:,1);
    ende = false;

    while(~ende)
        nr = nr+1;
        ph =  X h  *th/( t h  *th);
        ph = ph/norm(ph);
        thnew = Xh*ph/( p h  *ph);
        prec = (thnew-th)   *(thnew-th);
        th = thnew;

        if prec <= tol^2
            ende = true;
        elseif it <= nr
            ende = true;
             isp.(  Iteration   stops without convergence )
        end
    end
    Xh = Xh-th* p h  ;
    T(:,h+1) = th;
    P(:,h+1) = ph;
    oldVar = currVar;
    currVar = sum(var(Xh));
    if h >= 1
        pcvar(h+1) = 100*( oldVar     currVar )/varTotal;
    end
    nr = 0;

end
  pcvar(1) = [];
end
```

Cargar espectros

```
    function [Spectra, Wavenumbers,SpectraTitle,SpectraComments]= LoadSpectra (fid)
fseek(fid,30,    bof    );
SpectraTitle = {char(nonzeros(fread(fid,255,    uint8    ))    )};
fseek(fid,564,    bof    );
Spectrum_Pts = fread(fid,1,    int32    );
fseek(fid,576,    bof    );
Max_Wavenum = fread(fid,1,    single    );
Min_Wavenum = fread(fid,1,    single    );

% The Wavenumber values are assumed to be linearly spaced between
% between the Min and Max values. The array needs to be flipped
% around to get the order lined up with the absorbance data.

Wavenumbers = flipud(linspace(Min_Wavenum,
    Max_Wavenum,Spectrum_Pts).    )    ;


% The starting byte location of the absorbance data is stored in the
% header. It immediately follows a flag value of 3:

Flag = 0;

fseek(fid,288,    bof    );

while Flag ~= 3
    Flag = fread(fid,1,    uint16    );
end

DataPosition = fread(fid,1,    uint16    )    ;
fseek(fid,DataPosition,    bof    );

Spectra = fread(fid,Spectrum_Pts,    single    );

% Same story goes for the Comments section with a flag of 4.
% The size of the section is the difference between the two.

Flag = 0;

fseek(fid,288,    bof    );

while Flag ~= 4
    Flag = fread(fid,1,    uint16    );
end
% % % % % % % % CommentPosition = fread(fid,1,    uint16    )    ;
SpectraComments = {char(nonzeros(fread(fid,(DataPosition-DataPosition),    uint8
    ))    )};
fclose(fid);
end
```

Distancia

```
        function d = distancia(g,m)
S = size(g);
d = 0;
for i = 1:S(2)
    d = d+(g(i)-m(i)).^2;
end
d = sqrt(d);
end
```