UNIVERSIDAD EUROPEA DE MADRID

ESCUELA DE ARQUITECTURA, INGENIERÍA Y DISEÑO

MÁSTER UNIVERSITARIO EN ANÁLISIS DE DATOS MASIVOS

TRABAJO FIN DE MÁSTER

Análisis del manuscrito II/1391 (2) de la Colección Gondomar con espectroscopía Raman y técnicas de Big Data

FRANCISCO JOSÉ GÓMEZ FERNÁNDEZ

Dirigido por

Nicolás Coca López

CURSO 2024 - 2025

TÍTULO: Análisis del manuscrito II/1391 (2) de la Colección Gondomar con espectroscopía Raman y técnicas de Big Data

AUTOR: FRANCISCO JOSÉ GÓMEZ FERNÁNDEZ

TITULACIÓN: MÁSTER UNIVERSITARIO EN ANÁLISIS DE DATOS MASIVOS

DIRECTOR DEL PROYECTO: Nicolás Coca López

FECHA: 7 de septiembre de 2025

A mis padres, por su constante apoyo y guía.

Agradecimientos

En primer lugar, quiero dar las gracias a María del Valle Ojeda, de la Università Ca'Foscari Venezia, y Santiago Sánchez, del IEM-CSIC, por confiar en mí y darme la oportunidad de colaborar con ellos en un proyecto tan enriquecedor y novedoso como este, que me ha permitido conocer cómo funciona un proyecto muldisciplinar con ramas tan, a priori, dispares como la física-química teórica, la filología y la ciencia de datos. Ambos han aportado sus puntos de vista expertos para que, finalmente, todo el trabajo acabara llegando a buen puerto.

Agradecer la labor de mi tutor, Nicolás Coca, cuya guía y aportaciones han sido absolutamente fundamentales en este proyecto para alcanzar los resultados que se muestran a lo largo de este documento.

No quiero dejar pasar la oportunidad de dar las gracias a mis padres, mi hermana y mi familia en general, que me han dado toda su ayuda para llevar todo lo mejor posible las veces en que parecía que este curso podía llegar a complicarse demasiado. Extiendo también este agradecimiento a mis amigos más cercanos como Manuel, Jota, Vivi y Ádam, que han estado también al pie del cañón cuando más necesario era.

Destacar la ardua tarea de mi gran amigo y compañero Ekaitz, que se ha empleado a fondo dándome una ayuda y visión adicionales que a la larga han sido claves para poder desarrollar este proyecto de la manera en que finalmente se ha hecho.

Gracias también a ti, Inés. Tu apoyo incondicional todo este tiempo ha servido para aportar calma y sentir un respaldo total en todo momento que ha sido un pilar fundamental para terminar sacando adelante no solo este TFM y los proyectos asociados a él sino todo el máster.

Por último pero por supuesto no menos importante, muchísimas gracias a mis compañeros de trabajo, Lucía Pereira y Juan Torrejón, ambos investigadores del Instituto de Ciencias del Patrimonio (Incipit-CSIC), por introducirme en este precioso mundo del estudio del Patrimonio e irme generando curiosidad para seguir ahondando con cada vez más ganas en él y descubriendo distintas líneas de trabajo, a cada cual más interesante.

RESUMEN

El estudio de manuscritos históricos plantea importantes desafíos técnicos, especialmente cuando se trata de ejemplares escritos únicamente con tinta ferrogálica, cuya homogeneidad dificulta la obtención de información diferenciada mediante espectroscopía Raman. En este contexto, el presente trabajo desarrolla un flujo de análisis automatizado y replicable para el tratamiento masivo de espectros, con el fin de optimizar tanto su preprocesamiento como su posterior interpretación.

El proyecto incluye la revisión de la bibliografía existente sobre tintas ferrogálicas y la aplicación de la espectroscopía Raman a materiales patrimoniales. Se han optimizado los principales pasos de preprocesado -suavizado, corrección de línea de base, normalización y eliminación de artefactos - y se ha diseñado una aplicación que permite la corrección manual de espectros individuales afectados por *spikes*. En la fase analítica, se han implementado métodos de reducción de dimensionalidad, con énfasis en la factorización en matrices no negativas (NMF), junto con técnicas de clustering como Gaussian Mixture Models (GMM), evaluando sus parámetros y rendimiento en este contexto específico.

Los resultados obtenidos demuestran que la combinación de espectroscopía Raman y técnicas de análisis de datos constituye una herramienta eficaz para el estudio de manuscritos, reduciendo drásticamente los tiempos de procesamiento y aportando un marco robusto y reproducible. Estas aportaciones abren nuevas perspectivas para la investigación interdisciplinar en humanidades digitales y conservación del patrimonio.

Palabras clave: Stefanelo Botarga · Tintas ferrogálicas · Spikes · Espectroscopía Raman · NMF · Clustering

ABSTRACT

The study of historical manuscripts poses significant technical challenges, particularly in cases written exclusively with iron gall ink, whose spectral homogeneity hinders the extraction of differentiated information by Raman spectroscopy. In this context, the present work develops an automated and replicable workflow for the large-scale processing of spectra, aiming to optimize both preprocessing and subsequent interpretation.

The project includes a review of the existing literature on iron gall inks and the application of Raman spectroscopy to heritage materials. The main preprocessing steps—smoothing, baseline correction, normalization, and artifact removal—were optimized, and an application was designed to enable the manual correction of individual spectra affected by *spikes*. In the analytical stage, dimensionality reduction methods were implemented, with emphasis on Non-negative Matrix Factorization (NMF), together with clustering techniques such as Gaussian Mixture Models (GMM), evaluating their parameters and performance in this specific context.

The results demonstrate that combining Raman spectroscopy with data analysis techniques provides an effective tool for the study of manuscripts, drastically reducing processing times and offering a robust and reproducible framework. These contributions open new perspectives for interdisciplinary research in the fields of digital humanities and cultural heritage conservation.

Keywords: Stefanelo Botarga · Iron Gall Inks · Spikes · Raman Spectroscopy · NMF · Clustering

Índice general

Índice de Figuras

Índice de Tablas

Capítulo 1. RESUMEN DEL PROYECTO

1.1 Contexto y justificación

En el contexto del estudio y la conservación del Patrimonio Histórico, el patrimonio documental constituye una faceta de gran relevancia, ya sea con iluminación o sin ella.

En este caso, se tratará un manuscrito sin iluminación. En este aspecto, se estudian las tintas ferrogálicas, omnipresentes en los documentos entre la Edad Media y el siglo XX. Estas tintas, compuestas principalmente por sales de hierro/iones de hierro (el sulfato ferroso parece que es el más habitual) y ácidos orgánicos, están sujetas a procesos de degradación química.

La presencia de iones férricos y sustancias ácidas promueve reacciones de oxidación y acidólisis que, a largo plazo, deterioran tanto la tinta como el soporte celulósico, comprometiendo la estabilidad del documento. La espectroscopía Raman es una técnica no invasiva que ofrece la posibilidad de estudiar la composición molecular de estas tintas. De este modo, es posible obtener información detallada sobre la composición química, el estado de conservación y los posibles productos de degradación, lo que contribuye tanto al diagnóstico como al seguimiento del deterioro, respetando la integridad del documento.

Este estudio responde a la demanda creciente de técnicas analíticas respetuosas con el patrimonio cultural, alineándose con líneas de investigación en conservación preventiva y caracterización de materiales históricos.

1.2 Planteamiento del problema

Estudios previos han empleado FTIR (**F**ourier **T**ransform **I**nfrared **S**pectroscopy) y técnicas microquímicas, pero estas presentan limitaciones en cuanto a resolución y especificidad. La espectroscopía Raman permite una caracterización más precisa de las tintas ferrogálicas, sin requerir contacto o preparación de muestra.

En el ámbito empresarial y cultural, este proyecto podría servir como apoyo a bibliotecas y archivos para conservación de manuscritos. Desde una perspectiva científico-técnica, contribuye a desarrollar metodologías optimizadas para usar esta técnica en objetivos complejos, como los manuscritos antiguos.

1.3 Objetivos del proyecto

El objetivo general del proyecto es desarrollar un protocolo basado en espectroscopía Raman para caracterizar tintas ferrogálicas, determinando su composición química y evaluando su estado de degradación. Los objetivos específicos del mismo son estos:

- Revisión bibliográfica sobre las características moleculares de las tintas ferrogálicas y sobre la aplicación de la espectroscopía Raman en el estudio de materiales históricos.
- Optimización y automatización de los principales pasos de preprocesado de espectros

Raman (suavizado, normalización, detección y eliminación de artefactos).
- Selección y evaluación de diferentes algoritmos de reducción de dimensionalidad y clustering, junto con la optimización de sus parámetros, con el fin de obtener representaciones fiables y fácilmente interpretables de los datos.
- Desarrollo de una aplicación para la corrección manual e interactiva de espectros individuales con presencia de spikes, pensada como herramienta de apoyo para investigadores.

1.4 Resultados obtenidos

En este trabajo se ha logrado automatizar el preprocesamiento de un gran volumen de espectros Raman y desarrollar una aplicación que permite la limpieza manual de espectros individuales de forma intuitiva. Asimismo, se ha implementado un flujo de análisis replicable que combina la descomposición en componentes mediante NMF con técnicas de clustering como GMM, obteniendo unos resultados que han resultado de gran utilidad para el equipo de filólogos con el que se estaba colaborando. La metodología aplicada ha demostrado ser capaz de manejar de forma eficiente espectros de tinta ferrogálica en un manuscrito de elevada complejidad técnica, lo que refuerza la validez de los resultados y abre nuevas posibilidades de investigación en el ámbito de la filología y las humanidades digitales.

1.5 Estructura de la memoria

Este documento empieza con un apartado dedicado a los antecedentes y el estado del arte, los últimos avances en investigación en este sentido y la relevancia que podría tener un estudio de este tipo.

En la sección dedicada a los objetivos se desarrollará en detalle la lista de objetivos generales y específicos para este trabajo.

La sección de Desarrollo constará de varias partes. La primera de ellas, que servirá para exponer conceptos necesarios, explicará los fundamentos de la espectroscopía Raman y por qué es una técnica tan potente. En otra de las partes se tratarán aspectos como el preprocesamiento o los métodos de reducción de dimensionalidad y clustering de datos que puedan resultar útiles para estos datos desde un punto de vista teórico. se podrá explicar el proceso de toma y tratamiento de datos para cada técnica utilizada, obteniendo diferentes métricas.

En el apartado de Discusión, se compararán los diferentes métodos, recurriendo a la opinión de expertos externos en algunos de los campos de trabajo que abarcará este estudio para tener una visión mucho más completa sobre qué forma de trabajar es la mejor y cuál de los resultados obtenidos es más consistente.

Por último, podrán desarrollarse unas conclusiones y explicaciones de posibles líneas de trabajo a futuro.

Capítulo 2. ANTECEDENTES Y ESTADO DEL ARTE

2.1 Estado del arte

En los últimos años, ha aumentado en gran manera la cantidad de estudios que utilizan la espectroscopía Raman para analizar manuscritos con y sin iluminación (decoración pictórica hecha a mano en un manuscrito). Por ejemplo, Vandenabeele et al[1] exponen los avances realizados en la década previa a 2007, demostrando como el uso de este tipo de espectroscopia ha ido al alza en esos años, siguiendo la tendencia hasta nuestros días.

En concreto, esta técnica se ha utilizado mucho para el estudio de manuscritos iluminados, con el fin de identificar pigmentos de todas las épocas. Hay estudios analizando pigmentos de todas las regiones del mundo, como persas [2] o europeos [3], probando la variabilidad de esta técnica de caracterización. Por ejemplo, las bondades de esta técnica se pueden ver en casos como el siguiente, descrito en un artículo [4] dedicado a estudiar pigmentos prehistóricos presentes en cuevas:

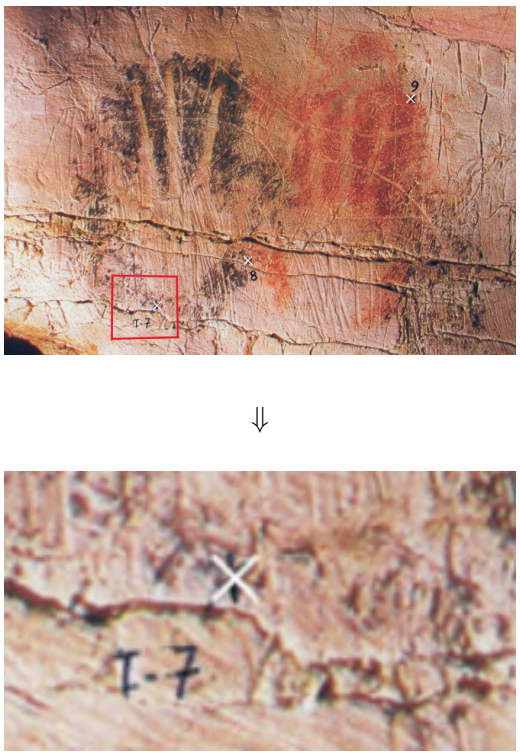

Figura 2.1. Zona analizada para pigmentos negros, referencia I-7.
[4]

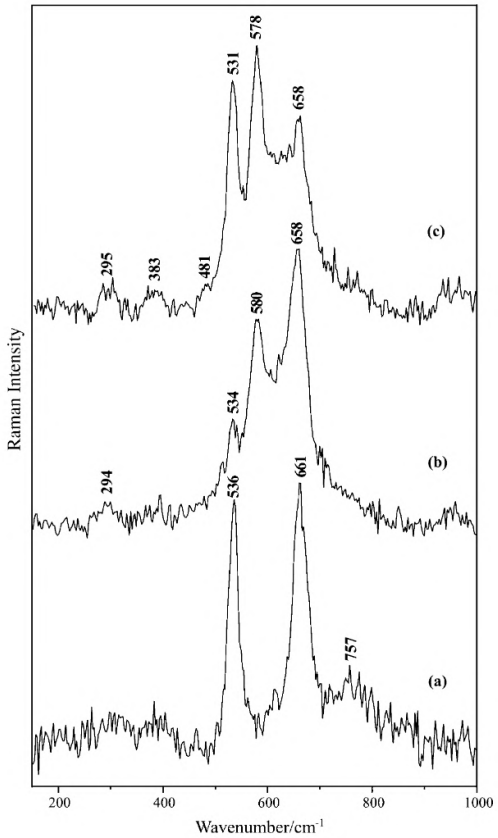

Figura 2.2. Espectro Raman de diferentes zonas de I-7 para pigmentos negros.
[4]

La Figura 2.2 muestra la potencia de esta técnica de caracterización a la hora de estudiar pigmentos que, a priori, tienen poca diferencia entre sí.

Esta cantidad de estudios habla de una importancia creciente del estudio del Patrimonio Histórico haciendo uso de la espectroscopía Raman.

No obstante, este método no ha tenido un gran desarrollo a la hora de investigar tintas del mismo tipo, puesto que en muchos casos las diferencias son indistinguibles con los métodos tradicionales de análisis de datos, que se utilizaban hasta hace pocos años. En este marco concreto, se tiene una serie de artículos que han ido investigando distintos componentes de las tintas ferrogálicas utilizando espectroscopía Raman:

- En un estudio de 2006, Lee et al. [5], se aplicó espectroscopía micro-Raman para estudiar el envejecimiento de las tintas ferrogálicas sobre pergamino. Se analizó la evolución química de la tinta en diferentes estados de conservación, llegando a la conclusión de que se podían distinguir componentes frescos de componentes envejecidos en virtud de pequeñas diferencias entre los espectros Raman. En concreto, se hizo especial hincapié en los principales compuestos que forman parte de estas tintas. Esto es, el ácido gálico, el hierro y los derivados de estos. No obstante, este estudio carece de un enfoque estadístico ni automatizado, pues los espectros eran analizados de forma cualitativa. El estudio resultó especialmente importante porque demostró que la tinta ferrogálica no es un compuesto químico inerte, sino que sufre una degradación medible.
- En un estudio de 2008, de nuevo Lee et al. [6] estudiaron las tintas ferrogálicas con espectroscopía Raman, tomando esta vez un enfoque mucho más completo, si bien se trata también de un estudio cualitativo. Se consideró un número de muestras mucho mayor que en el caso anterior, contando con muestras modernas y envejecidas para mejor comparación. Se evaluó con detalle la capacidad de la espectroscopía Raman para identificar los componentes de la tinta. Además de esto, se buscó también una serie de limitaciones clave que puede tener esta técnica para estudiar tintas de este tipo. Por ejemplo:
  - Fluorescencia del soporte o inducida por el propio láser utilizado.
  - Debilitamiento de la señal debido a la degradación del sustrato donde está la tinta.
  - Dificultad en la identificación de componentes clave, como el sulfato de hierro.

  Aunque se trató de un enfoque más completo, el análisis seguía basado en la inspección visual de los espectros, sin implementación de modelos estadísticos ni técnicas de minería de datos.
- En un nuevo estudio de 2022, Espina et al. [7] aplican un método más avanzado para analizar las tintas ferrogálicas. Se incluyeron técnicas como la espectroscopía Raman, espectroscopía infrarroja por transformada de Fourier (FTIR), espectroscopía UV-VIS o microscopía electrónica (SEM-EDX). Se estudió el envejecimiento de estas tintas tanto en manuscritos reales como en muestras hechas en laboratorio. Se puso mucho énfasis en la interacción entre la tinta y el sustrato, estudiando cómo estas cambian y se degradan con el paso del tiempo.
  En concreto, la espectroscopía Raman se aplicó con mucho éxito para identificar productos de oxidación y componentes orgánicos residuales. No obstante, si bien este estudio usa métodos más avanzados y rigurosos que los dos anteriores, no se aplica un procesamiento de grandes volúmenes de datos, operando solo con espectros pun-

tuales.

Gracias a un trabajo reciente de Vassiou et al. [8] se ha conseguido caracterizar la tinta ferrogálica sin envejecer, cuyos espectros son estos:

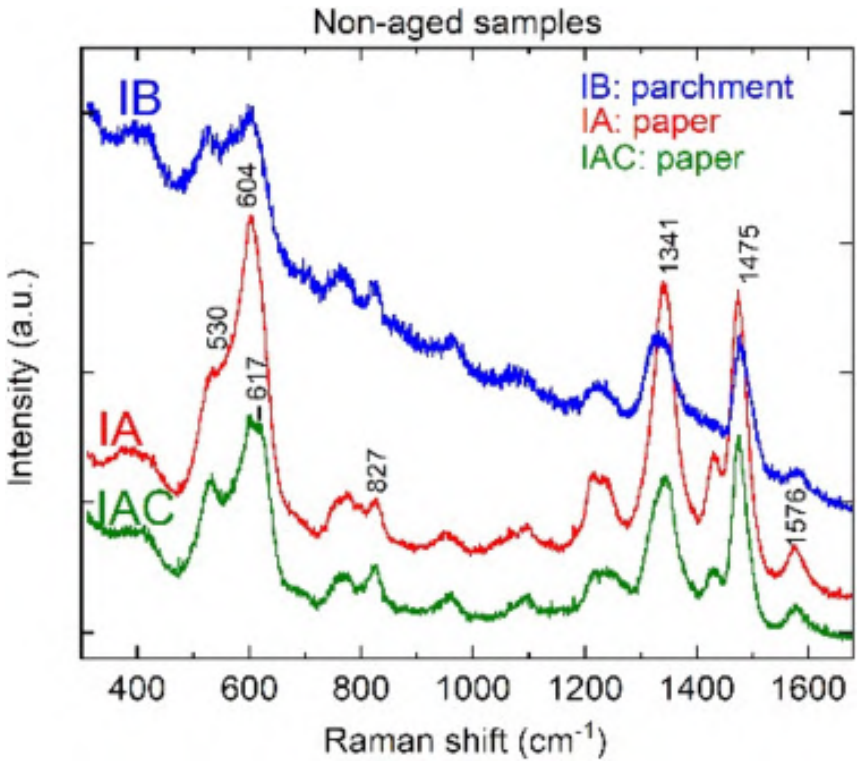

Figura 2.3. Espectros Raman de tintas ferrogálicas sin envejecimiento. [8]

En estos espectros se pueden detectar las siguientes bandas, caracterizadas por el artículo de Espina et al. [7], mencionado anteriormente:

Posición $[cm^{-1}]$	**Asignación vibracional**
$550 - 600$	Elongación del enlace Fe–O en complejos de hierro con taninos
$815 - 817$	Vibración del anillo en ácido gálico
$980 - 985$	Elongación simétrica en aniones sulfato en sulfato de hierro
$1005 - 1008$	Elongación simétrica en aniones sulfato en sulfato de calcio
$1395 - 1405$	Vibración de grupos carboxílicos en ácido gálico
$1420 - 1440$	Elongación de enlaces C–OH y vibraciones del anillo aromático
1430	Vibración asociada al enlace éster en taninos
$1477 - 1478$	Vibración del semianillo aromático de la unidad de ácido gálico
$1570 - 1590$	Elongación del anillo aromático de la unidad de ácido gálico

Tabla 2.1. Bandas Raman características de tintas ferrogálicas preparadas con recetas históricas. [7]

2.1.1 Reducción de dimensionalidad

Al tener un gran volumen de datos, será muy importante el intentar reducir la dimensionalidad de los mismos. No hacerlo podría llevar a dificultad para la convergencia de resultados por la complejidad de los modelos y el volumen de datos. Para ello, se hará uso de una amplia variedad de métodos, como los siguientes:

- **Principal Component Analysis (PCA)**: es un algoritmo que transforma un conjunto de variables en una serie de coordenadas independientes (componentes). El objetivo de cada una de estas componentes es explicar la máxima varianza posible del dataset.

  En un artículo de 2014, Bro et al. [9] explican en detalle este método, exponiendo entre otras cosas las fórmulas matemáticas que definen el algoritmo:

  $$X = T \cdot P^t + E \tag{2.1}$$

  En esta ecuación, se tiene:

  - $X \equiv$ matriz con los datos originales de los espectros.
  - $T \equiv$ matriz de scores $\longrightarrow$ representa las coordenadas de cada espectro en el espacio PCA.
  - $P^t \equiv$ matriz de loadings traspuesta $\longrightarrow$ describe el peso de cada variable original en las componentes.
  - $E \equiv$ matriz de residuos $\longrightarrow$ contiene la parte no explicada por los componentes impuestos.

  De la misma manera, también se define la ecuación de las componentes principales, que son los vectores propios (p_i) de la covarianza de los datos originales (X):

  $$S = \frac{1}{n-1} X^t \cdot X \tag{2.2}$$

  En esta ecuación, $n \equiv$ número de espectros contenidos en X.

  $$S \cdot p_i = \lambda_i \cdot p_i \tag{2.3}$$

  En esta ecuación, i hace referencia al número de componentes impuestas al algoritmo.

  El artículo habla también de la importancia que esta técnica puede tener para estudios de química y física, como es este caso.

  Para determinar el número de componentes más apropiado, se buscará determinar un umbral de varianza explicada a partir del cual añadir más componentes no aporta una información significativa. Típicamente, este umbral se define entre el 90 y el $95\,\%$.

- **Non Negative Matrix Factorization (NMF)**: es una técnica que descompone matrices imponiendo una restricción de no negatividad, lo que da lugar a representaciones de partes que son interpretables de forma inmediata si se tiene el conocimiento necesario. Esto es algo que no se da en otros métodos. En un artículo de 2013 [10], Wang et al.

estudian de forma muy exhaustiva las distintas variantes de este algoritmo, indicando también que es un método muy apropiado para quimiometría y análisis espectral, procesamiento de señales y disciplinas asociadas. La ecuación que define el NMF es la siguiente:

$$V \approx W \cdot H \tag{2.4}$$

En esta ecuación:

- $V \equiv$ matriz original, de dimensiones $m \times n$, que pueden ser en este caso espectros con m valores para $\Delta\omega$ y n espectros.
- $W \geq 0 \equiv$ matriz de componentes (loadings).
- $H \geq 0 \equiv$ matriz de coeficientes $\longrightarrow$ indica la contribución de cada componente.

Tanto W como H cumplen una ecuación que optimiza la distancia de Frobenius, siendo esta la optimización buscada:

$$\min_{\mathbf{W},\mathbf{H}\geq 0} \|\mathbf{V} - \mathbf{WH}\|_F^2 \tag{2.5}$$

Este algoritmo tiene, además, varios métodos de inicialización, pero en este caso se utilizará ***nndsvda***, que es el más apropiado para datos espectrales porque mejora la velocidad de convergencia y no induce ruido aleatorio, como sí hacen otros métodos.

Para determinar en este método el número de componentes a seleccionar, hay que buscar el número de componentes a partir del cual añadir más no reduce de forma significativa la distancia de Frobenius. También, dada la capacidad de este algoritmo para detectar componentes reales de espectros, es una buena opción comprobar las gráficas generadas en cada caso para determinar qué número de componentes devuelve unos loadings esperables para el contexto dado, como puede ser el espectro Raman del hierro, el del ácido gálico, el del carbón...

- **t-distributed Stochastic Neighbor Embedding (t-SNE)**: es un método basado en el cálculo de distancias con **perplexity**, que será el parámetro que acabará, en última instancia, definiendo la distribución de datos y el tiempo de ejecución. En artículo de Melit Devassy et al. [11] de 2020 muestra que este es un método que funciona muy bien con datos de espectroscopía. El estudio demuestra que se obtiene un espacio más fácilmente segmentable que el resultante de, por ejemplo, el PCA, si bien el tiempo de ejecución es mayor.

  En este caso no hay una definición matemática concreta como para los casos previos, sino que se tiene una distribución de similitudes (p_{ij}) para el espacio original. También se define un espacio reducido, que en el caso de este proyecto será bidimensional, donde se define una distribución q_{ij}. El objetivo del modelo será reducir la divergencia KL (Kullbak-Leibler) entre ambas distribuciones.

De esta manera, para cada par de puntos x_i y x_j:

$$p_{j|i} = \frac{e^{-\frac{(x_i^2 - 2x_i x_j + x_j^2)}{2\sigma_i^2}}}{\sum_{k \neq i} e^{-\frac{(x_i^2 - 2x_i x_k + x_k^2)}{2\sigma_i^2}}} \tag{2.6}$$

Ahora, al simetrizar:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n} \tag{2.7}$$

Ahora, para el espacio reducido:

$$q_{ij} = \frac{\left(1 + y_i^2 - 2y_i y_j + y_j^2\right)^{-1}}{\sum_{k \neq l} \left(1 + y_k^2 - 2y_k y_l + y_l^2\right)} \tag{2.8}$$

Teniendo las distribuciones ya, se puede determinar la función de costo a minimizar:

$$C = KL(P||Q) = \sum_{i \neq j} p_{ij} \cdot log\left(\frac{p_{ij}}{q_{ij}}\right) \tag{2.9}$$

Esta función de costo castigará la separación de vecinos cercanos, lo cual ayuda a mantener estructuras a nivel local.

- **Uniform Manifold Approximation and Projection (UMAP)**: esta técnica guarda muchas similitudes con t-SNE, pero se basa en principios diferentes. UMAP asume que los datos se distribuyen sobre una variedad (manifold) de menor dimensión, y su objetivo es preservar tanto la estructura local como la global del espacio original. Para ello, construye un grafo de vecinos en el espacio original mediante una función de conectividad ajustada con parámetros locales, y luego lo proyecta en un espacio de menor dimensión minimizando una función de coste basada en la entropía cruzada.

  De forma más técnica, primero se calcula la conectividad de los datos en el espacio original mediante una fórmula del tipo:

$$\mu_{ij} = 1 - e^{\left(-\left(\frac{d(x_i, x_j) - \rho_i}{\sigma_i}\right)\right)} \tag{2.10}$$

  En esta ecuación, ρ_i representa la distancia al vecino más cercano de x_i, y σ_i se ajusta para preservar la densidad local.

  En el espacio reducido, se utiliza una función de tipo sigmoidal para calcular la proximidad:

$$f(y_i, y_j) = \frac{1}{1 + a \cdot \|y_i - y_j\|^{2b}} \tag{2.11}$$

  siendo a y b parámetros que controlan la forma de la función.

Finalmente, la función de coste que se minimiza es:

$$\mathcal{L} = \sum_{(i,j)} \mu_{ij} \log(f(y_i, y_j)) + (1 - \mu_{ij}) \log(1 - f(y_i, y_j)) \tag{2.12}$$

$$\mu_{ij} = 1 - e^{\left(-\left(\frac{d(x_i,x_j)-\rho_i}{\sigma_i}\right)\right)} \tag{2.13}$$

donde ρ_i representa la distancia al vecino más cercano de x_i, y σ_i se ajusta para preservar la densidad local.

En el espacio reducido, se utiliza una función de tipo sigmoidal para calcular la proximidad:

$$f(y_i, y_j) = \frac{1}{1 + a \cdot \|y_i - y_j\|^{2b}} \tag{2.14}$$

siendo a y b parámetros que controlan la forma de la función. Finalmente, la función de coste que se minimiza es:

$$\mathcal{L} = \sum_{(i,j)} \mu_{ij} \log(f(y_i, y_j)) + (1 - \mu_{ij}) \log(1 - f(y_i, y_j)) \tag{2.15}$$

Al ser un método novedoso y muy potente, se están llevando a cabo muchos estudios en múltiples campos usando esta técnica. Por ejemplo, un trabajo de 2021, escrito por Sainburg et al. [12] estudia cómo hacer modelos de aprendizaje automático semisupervisado en base a la reducción de dimensionalidad que hace UMAP. Otro estudio más reciente, de 2024, redactado por Chen et al. [13], aborda cómo se podría trabajar para mejorar la clasificación de espectros Raman utilizando la técnica de UMAP para reducir la dimensionalidad y posteriormente hacer clustering del espacio UMAP resultante.

En este caso, para alcanzar el mejor resultado de este método habrá que optimizar dos parámetros [14]. Si bien el modelo tiene muchos parámetros más, los más importantes para un buen resultado son aquellos referidos a las características de los vecindarios o agrupaciones formados:

- n_neighbors: define el tamaño del vecindario.
- min_dist: controla lo compactos que son los vecindarios formados.

Como se puede comprobar, todos los métodos pueden llegar a ser de gran utilidad en este contexto y marco de investigación. Será muy ilustrativo comprobar cuál funciona mejor en combinación con los posteriores métodos de clustering.

2.1.2 Segmentación (clustering) de los datos

Se han empleado diferentes métodos de segmentación de datos para clasificar de forma no supervisada los datos que han sido sometidos a la reducción de dimensionalidad. Todos ellos operan de forma totalmente distinta y tienen características y funcionalidades diferentes. Estos son los métodos seleccionados:

- **K-Means:** es uno de los métodos de clustering no supervisado más utilizados por su

simplificidad y eficacia. Su objetivo es segmentar un conjunto de datos en k clusters, con k previamente definido, con el objetivo de minimizar la variación interna de cada uno de estos grupos.

Se asigna cada punto del dataset a un cluster cuyo centro (llamado **centroide**) esté más cercano. Estos centroides y distribuciones son recalculados hasta que la solución acaba convergiendo.

Este algoritmo se introdujo en 1967 con un artículo de James MacQueen [15]. Ha sido muy adoptado en campos como la minería de datos y el análisis espectral.

A pesar de las ventajas que tiene, también presenta algunos inconvenientes, como el hecho de que se requiere seleccionar un número de clusters (k) previamente fijado y que no es un método de clustering especialmente robusto frente a los outliers y la existencia de clusters no esféricos o con densidades muy distintas.

- **Gaussian Mixture Model (GMM):** es una generalización probabilística del K-Means: en vez de asignar cada punto a un único cluster, lo que hace es asignar unas probabilidades de que cada punto pertenezca a cada cluster. Cada uno de los clusters es una distribución normal multivariante y el modelo será una mezcla ponderada de todas estas distribuciones. De esta manera, la ecuación que define el GMM es esta:

$$p(x) = \sum_{k=1}^{K} \pi_k \cdot N\left(x|\mu_k, \sum_k\right) \tag{2.16}$$

  En esta ecuación, se cuenta con estos términos:

  - π_k es la probabilidad del cluster k, siendo el sumatorio de todos los $\pi_k = 1$.
  - $N\left(x|\mu_k, \sum_k\right) \equiv$ es la densidad gaussiana con media μ_k y covarianza $\sum_k$.
  - $x \equiv$ un registro del dataset.

  Este método es más flexible que K-Means porque permite clusters de forma elíptica, luego no tienen que ser necesariamente esféricos como en K-Means, y con diferentes densidades y varianzas. Esto genera clusters con formas no tan simples como los generados por K-Means, lo que puede generar resultados más sólidos en marcos como el de este trabajo, en los que la mayoría de los espectros no presentan grandes diferencias entre sí. Uno de los trabajos previos que hablan de su utilidad para la ciencia de datos y el aprendizaje automático fue un artículo de 2006 [16] donde se habla de cómo este algoritmo actúa en datos que presentan patrones comunes.

  En este trabajo, se utilizará para comprobar si el modelado probabilístico que ofrece obtiene una agrupación de espectros mejor que enfoques más estrictos o basados estrictamente en densidad.

- **Spectral:** este método está basado en un enfoque diferente a los algoritmos anteriormente mencionados, puesto que no se basa en optimizar distancias en el espacio de

los datos iniciales, sino en transformar el propio espacio de estos datos usando álgebra lineal.

Este método destaca especialmente al detectar clusters no convexos y con estructuras aparentemente arbitrarias. Esto puede hacer que esta técnica sea muy buena opción para los datos que se manejarán en este estudio. El procedimiento que sigue el algoritmo es explicado en un trabajo de 2001 [17] y es el siguiente:

1. **Construcción del grafo de similitud:** cada nodo del grafo G representa a un espectro y las aristas se determinan por similitud entre pares. Estas similitudes pueden calcularse de diferentes maneras, pero se utilizará un kernel gaussiano, cuya ecuación es esta:
$$w_{ij} = e^{-\frac{x_i^2 - 2x_i x_j + x_j^2}{2\sigma^2}} \tag{2.17}$$
2. **Construcción de la matriz Laplaciana:** mediante la ecuación
$$\mathcal{L} = \mathcal{D} - \mathcal{W} \tag{2.18}$$
En esta ecuación, se tiene:
   - $\mathcal{D} = \sum_j w_{ij} \equiv$ matriz de grados.
   - $\mathcal{W} \equiv$ matriz de similitudes, siendo cada término de esta todos los w_{ij} definidos antes.
3. **Cálculo de los autovectores:** se calculan los k vectores propios, asociados a los menores valores propios y construyendo la nueva representación de los datos en un espacio de menor dimensión para separar los clusters más fácilmente.
4. **Aplicación de K-Means:** sobre el espacio generado por los pasos anteriores como herramienta final para segmentación de los datos.

Este esquema explica el funcionamiento del método:

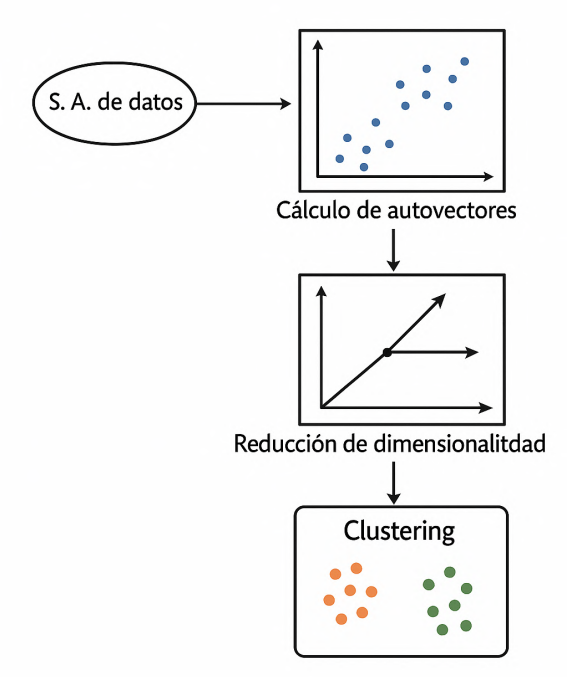

Figura 2.4. Funcionamiento de Spectral.
Imagen generada con Inteligencia Artificial.

El enfoque híbrido basado en grafos que aporta puede también resultar de gran utilidad para datos tan complejos como los de espectros Raman.

- **Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH):** este es un algoritmo de clustering jerárquico optimizado para operar con grandes conjuntos de datos. Funciona incrementalmente, pudiendo usar una memoria relativamente limitada y siendo muy eficiente con datasets grandes. Fue propuesto por Zhang et al. [18] en 1996 y ahora, con las nuevas técnicas de Big Data, se le está dando mucho uso.

  Se basa en la construcción incremental de una estructura jerárquica denominada **CF Tree** (Clustering Feature Tree), que resume el conjunto de datos mediante subclusters representados por tres estadísticas: el número de puntos contenidos (N), la suma de los vectores (L) y la suma de los cuadrados (SS). Esto permite calcular los centroides ($\mu = \frac{LS}{N}$) y el radio de los clusters y la dispersión dentro de los mismos. Este enfoque permite realizar una compresión efectiva de la información contenida en grandes datasets, reduciendo así la carga computacional de forma significativa.

El procedimiento seguido por el algoritmo se puede resumir de esta manera:

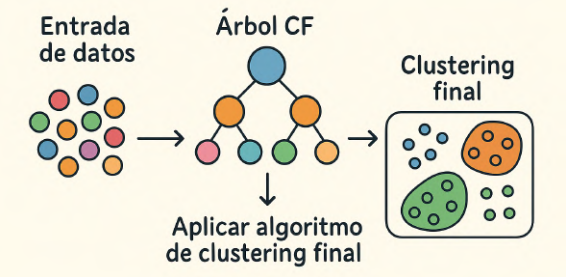

Figura 2.5. Explicación gráfica del funcionamiento de BIRCH. Imagen generada con Inteligencia Artificial.

En la versión implementada en este trabajo se utilizó un **umbral (*threshold*) fijo** como parámetro de control de compacidad para los nodos del árbol. Una vez construido el CF Tree a partir de dicho umbral, se realizó un proceso de búsqueda del número óptimo de clusters aplicando un algoritmo como K-Means sobre los centroides de los subclusters generados. Esto permite separar la fase de compresión (propia de BIRCH) de la fase de partición definitiva, haciendo el método más flexible y eficiente.

Esta técnica ha demostrado ser especialmente útil cuando se trabaja con grandes cantidades de datos complejos, ya que permite reducir la dimensionalidad y el número de observaciones de forma controlada, manteniendo la coherencia entre agrupaciones sin perder el carácter local de los datos.

- **Agglomerative Clustering:** es una técnica jerárquica de agrupamiento que sigue una estrategia ascendente (*bottom-up*), donde inicialmente cada punto es considerado como un cluster individual y, en cada iteración, se fusionan los dos clusters más cercanos hasta alcanzar un número determinado de grupos. Este enfoque permite construir un dendrograma que refleja la estructura jerárquica de los datos y puede adaptarse a diferentes criterios de fusión mediante el parámetro *linkage*.

  Se tienen diferentes métodos para linkage:

  - **linkage = *ward***: minimiza la varianza intra-cluster al fusionar aquellos grupos cuya unión produce el menor incremento en la suma total de cuadrados (*inercia*), lo que resulta especialmente útil para datos continuos y distribuciones esféricas.
  - **linkage = *complete***: considera como distancia entre dos clusters la distancia máxima entre cualquier par de puntos pertenecientes a dichos clusters. Esta variante es más sensible a la forma global de los grupos y tiende a producir agrupamientos más compactos y con mayor separación entre ellos.
  - **linkage = *single***: mide la distancia mínima entre los puntos de dos clusters. Es un método que tiende a formar cadenas de puntos conectados (efecto *chaining*), lo cual puede ser problemático si hay ruido o outliers en los datos.
  - **linkage = *average***: toma como referencia la distancia media entre todos los pares de puntos entre dos clusters. Es una opción intermedia entre las estrategias *single*

y *complete*, y ofrece resultados más equilibrados en muchos escenarios.

- **linkage = *centroid***: calcula la distancia entre los centroides (medias) de los clusters. Este enfoque puede generar resultados menos intuitivos, ya que el centroide de un grupo fusionado no tiene por qué coincidir con un punto real de los datos, y además no garantiza la monotonía de las distancias.

La diferencia en términos más visuales entre todos estos linkages se puede ver en esta gráfica, hechos todos los ejemplos sobre el mismo conjunto de datos:

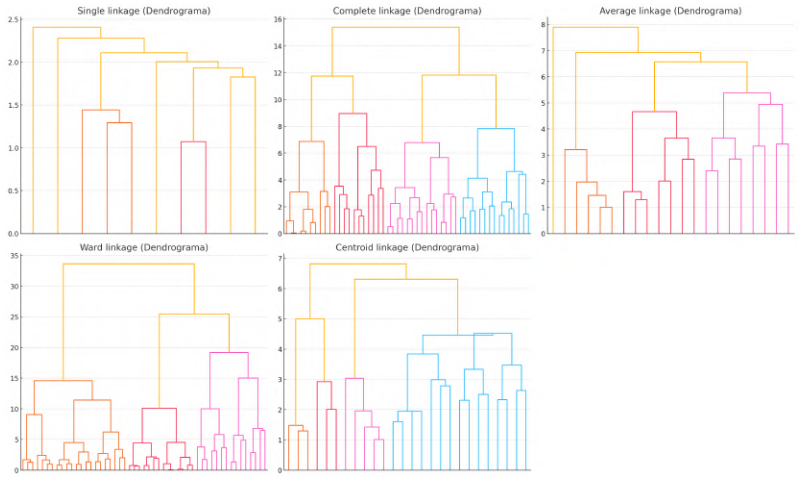

Figura 2.6. Diferencia entre los distintos linkages.
Elaboración propia.

Ambos métodos han sido aplicados tras una reducción de dimensionalidad previa, lo cual mejora la eficiencia computacional y permite obtener una segmentación más clara del espacio de datos. La comparación entre los resultados generados por ambos criterios de linkage permite identificar estructuras jerárquicas relevantes en los espectros Raman, diferenciando zonas con firmas químicas similares pero sutilmente distintas.

- **AffinityPropagation:** este algoritmo, propuesto por Frey y Dueck en 2007 [19], es una técnica de clustering basada en el intercambio de mensajes entre puntos de datos, sin necesidad de especificar a priori el número de clusters. En lugar de asignar arbitrariamente centroides, este método elige como **exemplars** (representantes del grupo) ciertos puntos del conjunto original.

  El algoritmo se basa en la actualización iterativa de dos matrices: la de *responsabilidad* ($r(i,k)$), que indica qué tan apropiado es que el punto k sea el ejemplar del punto i; y la de *disponibilidad* ($a(i,k)$), que refleja la idoneidad de que el punto i elija a k como su ejemplar teniendo en cuenta otros candidatos. Ambas matrices se actualizan con las siguientes expresiones:

$$r(i,k) \leftarrow s(i,k) - \max_{k' \neq k} \{a(i,k') + s(i,k')\} \tag{2.19}$$

  La flecha a la izquierda quiere decir que el valor de $r(i,j)$ se actualiza con el valor

calculado al lado derecho.

$$a(i,k) \leftarrow \text{mín}\left(0, r(k,k) + \sum_{i' \notin \{i,k\}} \text{máx}(0, r(i',k))\right) \tag{2.20}$$

$$a(k,k) \leftarrow \sum_{i' \neq k} \text{máx}(0, r(i',k)) \tag{2.21}$$

Donde $s(i,k)$ es la **similitud** entre los puntos i y k, que suele definirse como la distancia negativa entre ellos ($s(i,k) = -||x_i - x_k||^2$). Además, el valor de $s(k,k)$ se interpreta como la **preferencia** de que el punto k sea elegido como ejemplar.

El algoritmo continúa iterando estas actualizaciones hasta alcanzar la convergencia. Los puntos que maximizan la suma $a(i,k) + r(i,k)$ son seleccionados como ejemplares finales, y cada punto se asigna al ejemplar que le maximiza dicha suma.

Affinity Propagation es especialmente útil en conjuntos de datos con geometrías complejas y permite detectar automáticamente el número óptimo de clusters, aunque su coste computacional puede ser elevado en grandes datasets.

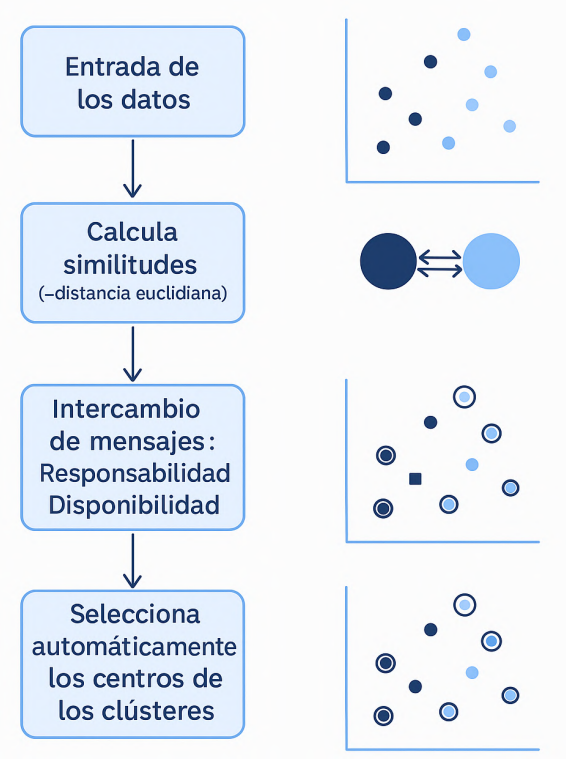

Figura 2.7. Procedimiento seguido por AffinityPropagation.
Imagen generada con Inteligencia Artificial.

Este método no debería funcionar bien por el tamaño del dataset, pero se puede comprobar cómo de sensible es el algoritmo a espectros sueltos muy característicos.

- **Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN):** es un algoritmo de clustering basado en densidad que extiende DBSCAN al generar una jerarquía de agrupamientos y seleccionar el nivel óptimo de corte mediante criterios de estabilidad. Esto permite una detección más robusta de clusters de forma arbitraria y densidades variables.

  En el artículo de McInnes et al. (2017) [20], se presenta este algoritmo como una mejora sustancial respecto a DBSCAN, especialmente en conjuntos de datos complejos como los espectrales, donde las densidades de los grupos pueden no ser homogéneas. El proceso puede resumirse en los siguientes pasos:

  1. Se construye un grafo ponderado donde los pesos entre puntos se basan en la **distancia mutua alcanzable**, que combina la distancia real entre puntos y la densidad local (medida como la distancia al k-ésimo vecino más cercano).

  2. Se genera un **árbol jerárquico** de clusters mediante un algoritmo de agrupamiento jerárquico (por ejemplo, single linkage).

  3. Se calcula la **estabilidad** de cada cluster (es decir, cuánto tiempo "vive" en la jerarquía) y se seleccionan los clusters más estables, eliminando aquellos poco consistentes o menos significativos.

  4. Los puntos que no pertenecen a ningún cluster suficientemente estable son etiquetados como **ruido**.

  El procedimiento se puede explicar con este esquema:

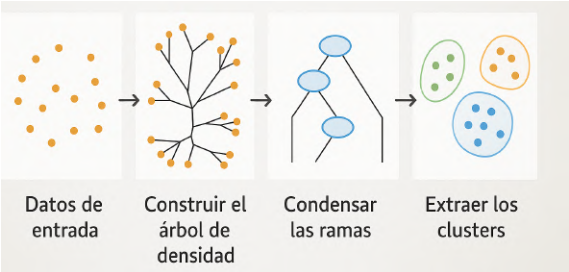

Figura 2.8. Funcionamiento de HDBSCAN.
Imagen generada con Inteligencia Artificial.

  Este algoritmo no requiere especificar el número de clusters, sino parámetros como *min_cluster_size*, que define el tamaño mínimo que debe tener un grupo para ser considerado cluster, lo cual es muy útil en tareas de exploración estructural. Además, es resistente al ruido y a formas no esféricas, siendo ideal para los espacios reducidos generados por PCA, t-SNE o UMAP.

- **Ordering Points To Identify Cluster Structure (OPTICS):** es un algoritmo de agrupamiento basado en densidad que, a diferencia de DBSCAN, no requiere definir un único

valor de ϵ (radio de vecindad) para identificar clusters. En lugar de eso, genera una estructura ordenada de los datos que permite visualizar agrupamientos a distintas escalas de densidad. OPTICS fue desarrollado para resolver las limitaciones de DBSCAN en escenarios con clusters de densidad variable. Tal como explican Ankerst et al. en un artículo de 1999 [21], este método se basa en dos conceptos:

- **Distancia de núcleo** (*core distance*): es la distancia desde un punto p hasta su k-ésimo vecino más cercano. Si un punto no tiene suficientes vecinos, su distancia de núcleo no está definida.
- **Distancia de acceso** (*reachability distance*): definida para un punto p con respecto a otro punto o, siendo la distancia a un k-ésimo vecino más cercano a o, con esta ecuación:

$$reachability\_dist(p, o) = max(core\_dist(o), d(p, o)) \tag{2.22}$$

El algoritmo genera una lista ordenada de puntos y sus respectivas distancias de acceso. A partir de esta lista se puede representar gráficamente la llamada **gráfica de accesibilidad** (*reachability plot*), donde los valles corresponden a agrupamientos y los picos a separaciones entre ellos. El flujo de trabajo del algoritmo es este:

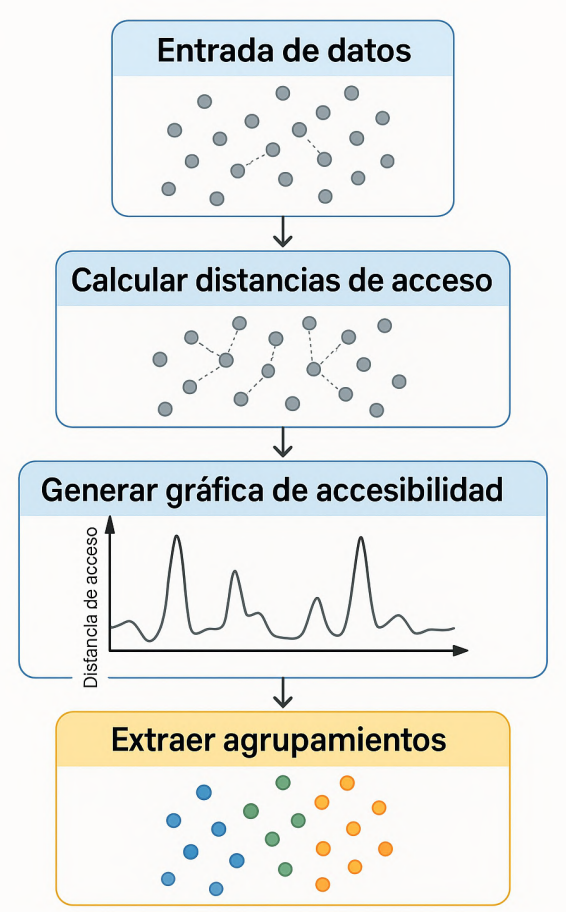

Figura 2.9. Procedimiento seguido por OPTICS. Imagen generada con Inteligencia Artificial.

OPTICS no devuelve directamente los clusters, sino una estructura que los sugiere. Se pueden extraer agrupamientos concretos cortando la gráfica de accesibilidad en diferentes niveles.
Este enfoque lo convierte en una excelente herramienta exploratoria para datasets espectrales, ya que permite detectar estructuras no esféricas, agrupamientos con formas complejas y clusters de diferentes densidades, como ocurre con los espectros Raman.

Para determinar los hiperparámetros de cada método de segmentación de datos se hará uso de cuatro métricas.

- **Inercia:** es la suma de las distancias al cuadrado entre cada punto del dataset y el centroide del cluster al que ha sido asignado [15]. Por tanto:

$$I = \sum_{i=1}^{k} \sum_{x \in C_i} ||x - \mu_i||^2 \tag{2.23}$$

En esta ecuación:

- $k \equiv$ número de clusters.
- $C_i \equiv$ puntos del cluster i.
- $\mu_i \equiv$ centroide del cluster i.

Cuanto menor es la inercia, más compacto serán los clusters generados. Este valor se utiliza para aplicar el método del codo. Este codo aparece en la gráfica para el número de clusters a partir del cual la mejora en esta métrica empieza a ser marginal.

- **Coeficiente de silueta:** es una métrica que sirve para determinar lo compactos que son los clusters formados [22]. Para cada punto x del dataset:

$$s = \frac{b - a}{max(a, b)} \tag{2.24}$$

En esta ecuación:

- $a \equiv$ media de distancias de cada punto x a los puntos del cluster al que está asignado.
- $b \equiv$ mínima media de distancias de x a todos los demás clusters.

Esta métrica toma valores $s \in [-1, 1]$. El objetivo será maximizar el promedio de todos los valores s.

- **Coeficiente de Calinski-Harabasz:** se utiliza para determinar la varianza entre clusters e intra cluster [23]. Viene dado por esta ecuación:

$$CH = \frac{Tr(B_k)}{Tr(W_k)} \cdot \frac{n - k}{k - 1} \tag{2.25}$$

En esta ecuación, se tienen estos términos:

- $n \equiv$ número total de datos.
- $k \equiv$ número de clusters.

- $Tr(B_k) \equiv$ traza de la matriz de dispersión entre clusters.
- $Tr(W_k) \equiv$ traza de la matriz de dispersión dentro de los clusters.

Este coeficiente debe maximizarse. Cuanto más alto sea, mejor separación entre clusters y menor dispersión interna se estarán manejando.

- **Índice de Davies-Bouldin:** mide la media de la peor similitud entre cada cluster y el más similar a este [24]. Se determina con esta ecuación:

$$DB = \frac{1}{k}\sum_{i=1}^{k} \max_{j \neq i} \left(\frac{s_i + s_j}{d_{ij}} \right) \tag{2.26}$$

En esta ecuación:

- $s_i \equiv$ media de distancias entre los puntos del cluster i y su centroide.
- $d_{ij} \equiv$ distancia entre centroides de los clusters i y j.

Este índice debe minimizarse, pues un menor valor indica que los clusters son más compactos y están más separados entre sí.

Tras esto, habrá que determinar el espectro medio de los clusters obtenidos para cada método tanto de reducción de dimensionalidad como de segmentación y estudiar la distribución de estos clusters para las páginas analizadas.

2.2 Recursos requeridos

Para la adquisición de datos se ha utilizado el **espectrofotómetro Raman Renishaw Virsa**. Para el posterior análisis de datos se ha hecho uso de diferentes librerías de Python para desarrollo de aplicaciones, visualización, segmentación y análisis de datos.

Para el **desarrollo de la aplicación** se utiliza la librería **Dash**.

Para la **visualización**, se utilizan las archiconocidas librerías **Matplotlib** y **Seaborn**, así como **Plotly** para visualizaciones interactivas.

Para el **análisis y procesamiento de datos** se utilizan librerías como **Scipy**, **Numpy**, **Pandas** y **Math**.

Para la **reducción de dimensionalidad y** el **clustering** se usan las librerías **hdbscan**, **umap**, **optics** y **scikit-learn** y, en especial y dentro de esta, los **módulos cluster**, **manifolda** (TSNE) y **decomposition** (PCA, NMF).

2.3 Contexto y justificación

Tal como se ha comentado en la Introducción, hay una gran cantidad de estudios que tratan la problemática de la conservación sin deterioro del Patrimonio Histórico documental con tintas ferrogálicas. Para estudiar su degradación en detalle, se han utilizado hasta ahora una gran variedad de técnicas, como FTIR [25][26], imágenes hiperespectrales [27] y

multiespectrales [28]. Estas técnicas han mostrado resultados variados dependiendo de las características de la tinta y el sustrato (papel, pergamino, roca...) estudiados.

El objetivo de este proyecto es aprovechar la mayor sensibilidad que presenta la espectroscopía Raman para detectar moléculas complejas, como las presentes en las tintas ferrogálicas. Esta técnica es sumamente útil para este tipo de estudios porque permite analizar de forma no invasiva los materiales de interés. Ha sido ampliamente utilizada para estudiar pigmentos de todo tipo, pero el potencial de la misma no ha sido totalmente explorado para estudiar las tintas ferrogálicas en el Patrimonio Histórico documental, lo cual deja abierta una puerta con grandes oportunidades para avanzar de gran manera en esta línea de investigación.

Las diferencias sutiles entre los espectros Raman de tintas que podrían parecer idénticos en una inspección visual pueden resultar fundamentales en el contexto de esta investigación, pues una pequeña diferencia entre componentes químicos de la tinta podría servir para explicar y prevenir una potencial degradación de los manuscritos y piezas de Patrimonio Histórico que tengan estas tintas tan comunes.

La gran sensibilidad que ofrece esta técnica de caracterización aplicada al estudio de manuscritos genera un nuevo reto: el análisis de datos masivos de una gran cantidad de espectros que presentan unas diferencias muy pequeñas entre sí. La revisión de estudios anteriores demuestra que estos no se han llevado a cabo de forma sistemática, sino con análisis visuales de espectros Raman o haciendo una selección manual de los mismos.

2.4 Planteamiento del problema

Este proyecto pretende explorar la vía de investigación abierta en este sentido. Se aplicará un tratamiento de datos a un conjunto de alrededor de 1000 espectros Raman (con mas de mil puntos por espectro) tomados de un manuscrito del siglo XVI de la **colección Gondomar**, preservado en la Real Biblioteca, localizada en el Palacio Real, en Madrid.

En este marco se podrá no solo analizar las tintas ferrogálicas en términos físico-químicos de evolución temporal y degradación, sino también el potencial de la espectroscopía Raman como herramienta para caracterización automatizada de materiales tan complejos como estas tintas. La creación de modelos de clustering y la identificación de componentes químicos en base a datos espectrales puede abrir nuevas puertas a la conservación preventiva, la restauración y el estudio histórico-científico verdaderamente disciplinar de Patrimonio Histórico.

Se buscará, por tanto, aportar una contribución metodológica sólida al campo de la conservación de patrimonio documental, uniendo la gran sensibilidad de la espectroscopía Raman con las capacidades derivadas del análisis avanzado de datos masivos (Big Data) para avanzar a herramientas más objetivas, reproducibles y escalables que las presentadas hasta la fecha.

Capítulo 3. OBJETIVOS

3.1 Objetivos generales

El objetivo general de este estudio es obtener un método optimizado y generalizable para analizar los espectros Raman de tintas ferrogálicas de manuscritos antiguos, permitiendo detectar las características principales de los mismos.

En concreto para el manuscrito problema, el objetivo principal será determinar cómo se ha desarrollado el génesis del mismo en base a las variaciones de los espectros Raman.

3.2 Objetivos específicos

Como objetivos en específico para conseguir el objetivo mencionado, se tienen los siguientes:

- Conseguir un procedimiento para localizar y eliminar los spikes presentes en los espectros.
- Hacer una aplicación que sirva para eliminar estos spikes en series de datos pequeñas o en las que el procedimiento anteriormente mencionado no funcione tan bien.
- Analizar los datos utilizando diferentes métodos de reducción de dimensionalidad.
- Obtener los componentes principales de las tintas ferrogálicas analizadas.
- Hacer una segmentación (clustering) de los datos para encontrar diferentes tipos de tintas en base a los componentes de las mismas, usando diferentes algoritmos y comparando los resultados.
- Utilizar los diferentes resultados obtenidos para hacer un mapa del génesis del manuscrito, en connivencia con el criterio de diferentes expertos de varias disciplinas.

3.3 Beneficios del proyecto

El uso de técnicas de análisis de Big Data pretende facilitar la detección de pequeñas variaciones en el espectro Raman de tintas ferrogálicas.

Hasta ahora, la mayoría de estudios con esta técnica se han centrado en pigmentos por su mayor variabilidad y la facilidad que ofrecen para detectar cambios. No obstante, la aplicación de métodos actuales de análisis de Big Data permite identificar variaciones mínimas en los espectros, hasta ahora indetectables con enfoques tradicionales. Esto abre la posibilidad de clasificar automáticamente distintos tipos de tinta y, en particular, de analizar de manera más completa las tintas ferrogálicas con procedimientos menos invasivos, lo que resultaría de gran utilidad para bibliotecas y archivos en el estudio y conservación de colecciones manuscritas.

Capítulo 4. DESARROLLO DEL PROYECTO

4.1 Planificación del proyecto

Los datos fueron tomados y cedidos por el Profesor de Investigación D. Santiago Sánchez Cortés, del IEM-CSIC, en el marco de un estudio de la Università Ca'Foscari Venezia sobre la obra de, Stefanello Botarga, un dramaturgo italiano del siglo XVI. Para tomar los datos, D. Santiago accedió a la Real Biblioteca entre dos y tres veces por semana durante casi dos años. La organización para llevar a cabo este proyecto, sin contar con el proceso de toma de datos, ha sido esta:

Actividad	Descripción	Duración (semanas)	Horas usadas
Estudio del estado del arte	Revisión detallada de estudios previos sobre espectroscopía Raman y tintas ferrogálicas	2	15
Estudio de alternativas actuales	Análisis de técnicas existentes (FTIR, HSI, MSI) y su aplicabilidad comparada	1	5
Análisis de informes previos	Estudio de documentación técnica sobre manuscritos y datos espectrales disponibles	1	10
Procesado de espectros Raman	Limpieza, normalización y organización del conjunto de espectros Raman	2	30
Diseño de sistema de análisis	Definición del pipeline para análisis masivo de datos: preprocesado, clustering y visualización	2	30
Desarrollo de scripts y algoritmos	Implementación en Python de modelos de reducción de dimensionalidad, clustering y análisis espectral	4	60
Diseño de validación y pruebas	Diseño de estrategia para comprobar estabilidad de los clusters y validez espectral	1	10
Evaluación de resultados	Interpretación de agrupamientos, elaboración de visualizaciones, comparación con conocimiento previo	2	25
Redacción del documento final	Escritura estructurada del TFM, incluyendo análisis, figuras y anexos técnicos	4	60
Reuniones y coordinación	Sesiones con tutores, especialistas del laboratorio y equipo del proyecto para toma de decisiones	4	10

Tabla 4.1. Cronograma de actividades realizadas en el proyecto.

4.2 Descripción de la solución, metodologías y herramientas empleadas

4.2.1 Adquisición de los datos

El trabajo de adquisición de los datos se realizó con el siguiente montaje:

Figura 4.1. Montaje experimental utilizado.

En la presente imagen, se pueden ver estos elementos:

a - Ordenador: el software con el que funciona este dispositivo (WiRE) es el propio que proporciona Renishaw, que tiene herramientas para adquisición de espectros, capturas de imagen instantáneas de las muestras, toma de espectros Raman en la forma de espectros puntuales, hacer mapeos hiperespectrales, etc. y para el procesamiento de datos, como eliminación de spikes, substracción de lineas base, etc.

b - Espectrofotómetro: se utilizó el **espectrofotómetro Raman Renishaw Virsa**, que tiene una fibra óptica de 5 metros para análisis a distancia de muestras y está acoplado a un

microscopio con 5, 20 y 50 aumentos, llegando a obtener una resolución de hasta $1\ \mu$m. Esto permite seleccionar las zonas de análisis de forma correcta.

c - Plomadas: se trata de una serie de fragmentos de plomo forrados con tela. Se utiliza para evitar que el calor inducido por el láser al papel acabe combando el mismo, lo que podría traducirse en una pérdida de calidad de los datos adquiridos o incluso en un daño irreparable en el manuscrito.

d - Manuscrito: contiene varios textos pertenecientes a la Colección Gondomar, una de las más importantes que se han conseguido conservar hasta nuestros días. En concreto, la parte que se analizará en este proyecto corresponde a las páginas que van de la 25 a la 66 y es atribuida a Stefanello Bottarga, un conocido dramaturgo veneciano [29] nacido alrededor de 1540.

Cuando un haz de luz monocromática (como el láser usado) incide sobre una muestra de estudio, los fotones pueden interactuar de diferentes maneras, habiendo tres formas de dispersión diferentes:

1. **Dispersión elástica**: también llamada **dispersión Rayleigh**. Es el fenómeno más dominante, dándose en aproximadamente el $99{,}9999\,\%$ de los casos.
   El fotón incidente interactúa con la nube electrónica de la molécula estudiada pero no se intercambia energía vibracional con la radiación. Consecuentemente, el fotón dispersado mantendrá la energía y frecuencia del incidente. Es necesario suprimir esta dispersión mediante filtros ópticos, pues rodea la señal Raman y complica mucho la adquisición de un buen espectro.

2. **Dispersión inelástica**: también llamada **dispersión Raman**. Este tipo de dispersión se da para aproximadamente uno de cada 10^6 de los fotones incidentes.
   En este nuevo caso, la interacción entre fotón y molécula acaba provocando un cambio en el estado vibracional de esta. El fotón dispersado presenta, por lo tanto, una energía diferente a la del fotón incidente. Se pueden dar dos situaciones:

   - **Dispersión Stokes**: el fotón transfiere parte de su energía a la molécula, por lo que sale dispersado con una menor energía, lo que reduce su frecuencia ($\nu < \nu_0$).
   - **Dispersión anti-Stokes**: se da cuando la molécula ya estaba vibracionalmente excitada. Cuando el fotón interactúa con ella, la molécula libera energía y el fotón es dispersado con una energía superior a la inicial, aumentando su frecuencia ($\nu > \nu_0$).

   Dentro de esto, la probabilidad de una dispersión Stokes es mucho más alta que la de una dispersión anti-Stokes, puesto que la mayoría de moléculas estarán en condiciones normales en su estado fundamental.

3. **Otros fenómenos**: se pueden dar otros fenómenos de interacción óptica como la **fluorescencia** o la **dispersión Brillouin**, que no son parte del efecto Raman pero pueden aparecer simultáneamente y han de ser tenidos en cuenta, pues en muchos casos van a enmascarar o deformar la señal Raman.

En esta figura, extraída de un trabajo de Otero et al. [30], se pueden ver las diferencias en cuanto a niveles energéticos vibracionales entre las diferentes formas en que se dispersa la luz al incidir sobre una muestra.

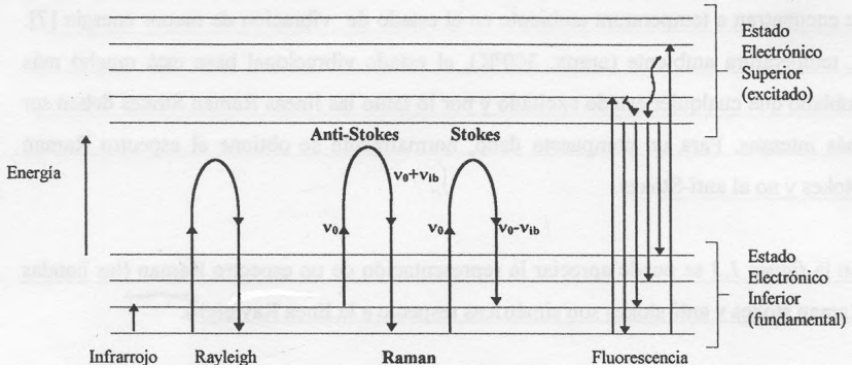

Figura 4.2. Dispersiones Rayleigh y Raman.
[30].

Esta otra figura sirve para ver de forma simplificada los tipos de dispersión de la luz cuando esta incide sobre una muestra:

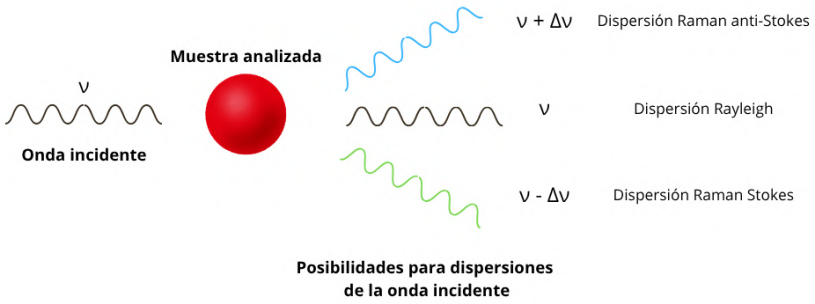

Figura 4.3. Dispersión Raman.
Elaboración propia.

El espectro Raman de la muestra analizada tiene dos variables:

- **Desplazamiento Raman** ($\Delta\omega$): es una comparación entre las longitudes de onda de las radiaciones incidente y dispersada. Se mide en cm^{-1} y se calcula mediante esta ecuación:

$$\nu = \frac{1}{\lambda} \longrightarrow \Delta\omega = \left(\frac{1}{\lambda_0} - \frac{1}{\lambda}\right) \cdot 10^7 \tag{4.1}$$

  Donde λ_0 es la longitud de onda del laser. Para este caso particular, $\lambda_0 = 785\ nm$. El factor 10^7 está en la ecuación porque la longitud de onda está en nm y el desplazamiento Raman está en cm^{-1}.

- **Intensidad Raman** (I): es el número de fotones dispersados para cada valor de $\Delta\omega$ que cubre el aparato de medida. La ecuación que define esta magnitud no es trivial, pero es proporcional a:

$$I \propto I_0 \cdot \left(\frac{1}{\lambda_0}\right)^4 \cdot \left|\frac{d\alpha}{dQ}\right|^2 \tag{4.2}$$

En esta ecuación, se tienen estos términos:

- $I \equiv$ intensidad de la señal Raman dispersada.
- $I_0 \equiv$ intensidad del láser incidente.
- $\lambda_0 \equiv$ longitud de onda del láser (en cm).
- $\frac{d\alpha}{dQ} \equiv$ derivada de la polarizabilidad (α) respecto de una coordenada normal vibracional (Q).

También se tiene la posibilidad de hacer un mapa Raman, para estudiar cómo se distribuyen los espectros Raman en una región del manuscrito. Esta funcionalidad permite obtener una imagen hiperespectral, en la que cada píxel contiene un espectro Raman completo. A partir de estos datos, es posible visualizar la localización de determinados compuestos, analizar zonas con comportamiento químico anómalo, o realizar segmentaciones espectrales para distinguir entre tintas, sustratos o zonas degradadas.

4.2.2 Preprocesamiento de los datos

Para hacer el preprocesamiento se siguen estos pasos:

1. **Lectura de todos los archivos de texto**, que constan de dos columnas: una para el desplazamiento Raman, medido en cm^{-1}, y otra para la intensidad Raman, medida en Cuentas (fotones que alcanzan el contador).

Función para leer los archivos txt

```python
def lectura_archivos(carpeta_principal):
    """
    Lee archivos .txt con columnas Wave e Intensity desde subcarpetas.

    Devuelve un DataFrame largo con columnas: Wave, Intensity, Archivo.
    """
    datos = []

    # Bucle para recorrer subcarpetas y archivos
    for subdir,_,files in os.walk(carpeta_principal):

        # Bucle anidado para cada archivo en la subcarpeta
        for file in files:

            # Filtrado para leer solo archivos .txt que no sean copias
            if file.endswith('.txt') and '- copia' not in file:
                filepath = os.path.join(subdir,file)           # Ruta completa del archivo
                try:
                    df = pd.read_csv(filepath,
                                     sep = r'\s+',             # Espacios en blanco como separadores
                                     comment = '#',            # Se ignoran líneas que comienzan con '#'
                                     names = ['Wave','Intensity'],  # Nombres de las columnas
                                     engine = 'python')        # Motor de parsing

                    # Conversión de tipos de datos y manejo de errores
                    df['Wave'] = pd.to_numeric(df['Wave'],errors = 'coerce').round(5)   # Redondeo a 5 decimales
                    df['Intensity'] = pd.to_numeric(df['Intensity'],errors = 'coerce')
                    df['Archivo'] = os.path.splitext(file)[0]
                    datos.append(df)
                except Exception as e:
                    print(f"Error leyendo {filepath}: {e}")

    return pd.concat(datos,ignore_index = True)
```

Figura 4.4. Función para leer todos los archivos .txt con datos.

2. **Interpolación lineal de los valores de desplazamiento Raman** con respecto al primer espectro medido, pues esta variable no ha tomado los mismos valores en todos los espectros y es necesario que tenga los mismos valores en todos los casos. Habrá, por tanto, que interpolar también los valores de intensidad Raman. El último paso es condensar todos los datos en una sola tabla. La función utilizada es esta:

Función para interpolar datos y unificarlos en un solo dataset

```python
def interpolar(df):
    """
    Interpola todos los espectros a la malla de desplazamientos Raman del primer espectro encontrado.

    Devuelve un DataFrame pivotado: filas = archivos, columnas = Wave.
    """
    # Extracción de desplazamientos Raman del primer espectro válido
    primer_nombre = df['Archivo'].unique()[0]
    wave_ref = df[df['Archivo'] == primer_nombre]['Wave'].dropna().values

    # Lista inicialmente vacía para los espectros interpolados
    df_interp_list = []

    # Iteración sobre cada espectro, agrupando por la columna Archivo
    for nombre,grupo in df.groupby('Archivo'):
        grupo = grupo.dropna()          # Eliminación de filas con NaN
        if len(grupo) < 2:
            continue

        # Interpolador lineal con relleno NaN fuera de rango
        f_interp = interp1d(grupo['Wave'],grupo['Intensity'],
                            kind = 'linear',
                            bounds_error = False,
                            fill_value = np.nan)
        intensidad_interp = f_interp(wave_ref)

        # Creación de fila como diccionario: clave = columna de longitud
        fila = {'Archivo': nombre}
        fila.update({round(w,5): i for w,i in zip(wave_ref,intensidad_interp)})
        df_interp_list.append(fila)

    # Unión de todas las filas en un DataFrame
    df_final = pd.DataFrame(df_interp_list)
    df_final.set_index('Archivo',inplace = True)
    return df_final
```

Figura 4.5. Función para interpolar todos los espectros a los mismos valores de desplazamiento Raman.

Se usará el formato de tabla pívot. De esta manera, cada línea corresponde a un único espectro, siendo el índice el nombre del archivo .txt, los nombres de las columnas serán los valores de desplazamiento Raman y los valores asociados a cada columna serán los de la intensidad Raman para cada punto.

3. **Suavizado de los espectros** mediante la aplicación de un filtro de media móvil con una ventana de 5 puntos. Este procedimiento consiste en sustituir el valor de cada punto de la señal por el promedio aritmético de los valores comprendidos en un intervalo (ventana) centrado en dicho punto. La ventana se desplaza secuencialmente a lo largo del espectro, recalculando en cada posición el valor medio correspondiente, lo que permite atenuar el ruido de alta frecuencia sin alterar de forma significativa la tendencia global de la señal.

Función para aplicar un filtro de media móvil

```python
def suavizado(df,window_size = 5):
    """
    Aplica una media móvil a cada espectro (fila) del DataFrame.
    window_size: número de puntos usados para calcular la media móvil.

    Devuelve un nuevo DataFrame con los espectros suavizados.
    """
    df_suavizado = df.apply(lambda fila: fila.rolling(window = window_size,          # Número de puntos en la ventana
                                                      min_periods = 1,               # Mínimo de puntos para calcular la media
                                                      center = True).mean(),axis = 1)  # Centrado de la ventana y promediado
    return df_suavizado
```

Figura 4.6. Función para suavizar los espectros.

4. **Eliminación de los spikes**, mediante un algoritmo expuesto por Nicolás Coca [31]. Estos spikes se generan cuando rayos cósmicos, partículas cargadas de alta energía, inciden en uno o varios pixels del detector. La apariencia de un spike es un pico muy alto y estrecho que aparece aleatoriamente en cualquier punto del espectro Raman.

   Este algoritmo se basa en la función *find_peaks* que ofrece *scipy*, ajustando parámetros de los picos como la prominencia, que es la altura mínima que hay que descender

desde un pico para encontrar el siguiente, o la altura relativa a la que se calcula el ancho de los mismos.

Aquí se puede ver cómo actúa una variación de prominencia para la detección:

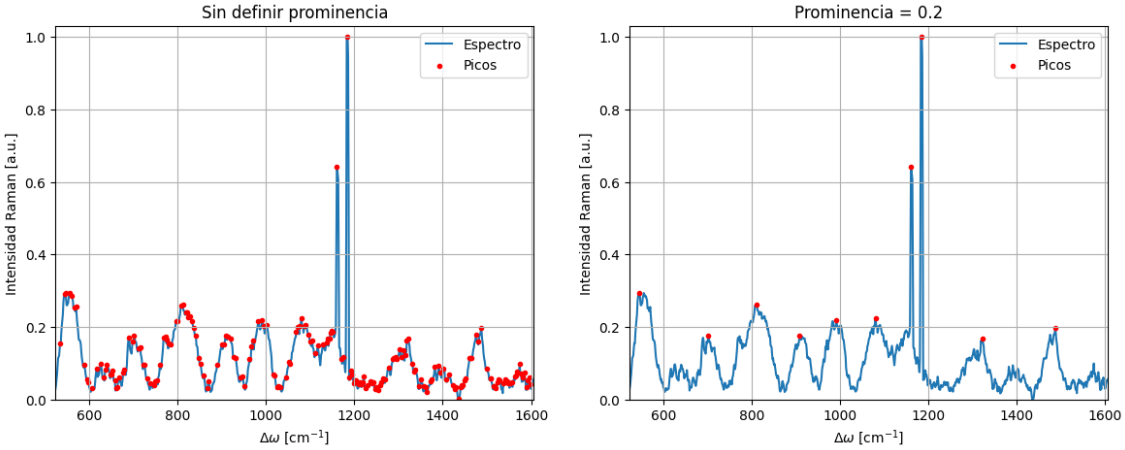

Figura 4.7. Influencia del parámetro de prominencia del pico.

Asimismo, se puede comprobar cómo funciona la detección del ancho de cada pico:

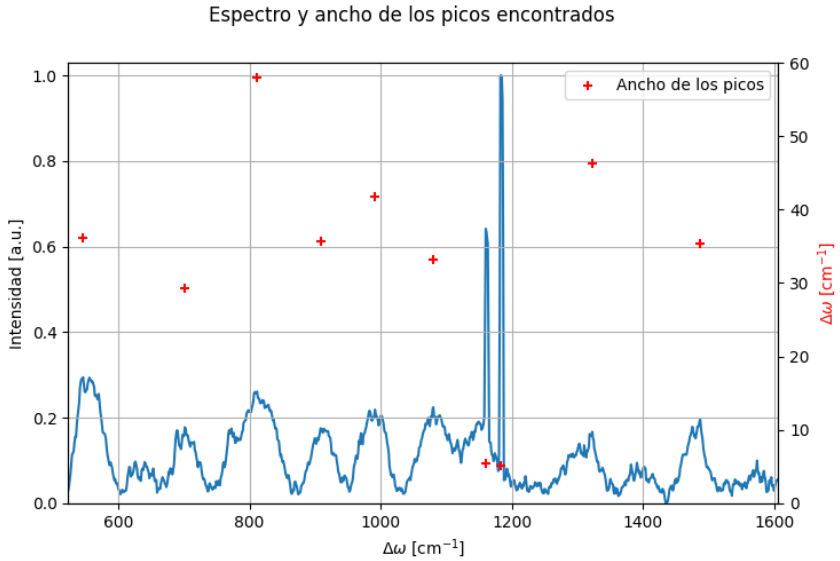

Figura 4.8. Ancho de los picos detectados.

Una vez se detecta un pico y se etiqueta como rayo cósmico (spike, a partir de ahora), se busca hacer una interpolación lineal en las cercanías del punto donde se encuentra dicho outlier.

Seleccionando 10 puntos alrededor de cada spike para la interpolación, se pueden obtener resultados como este:

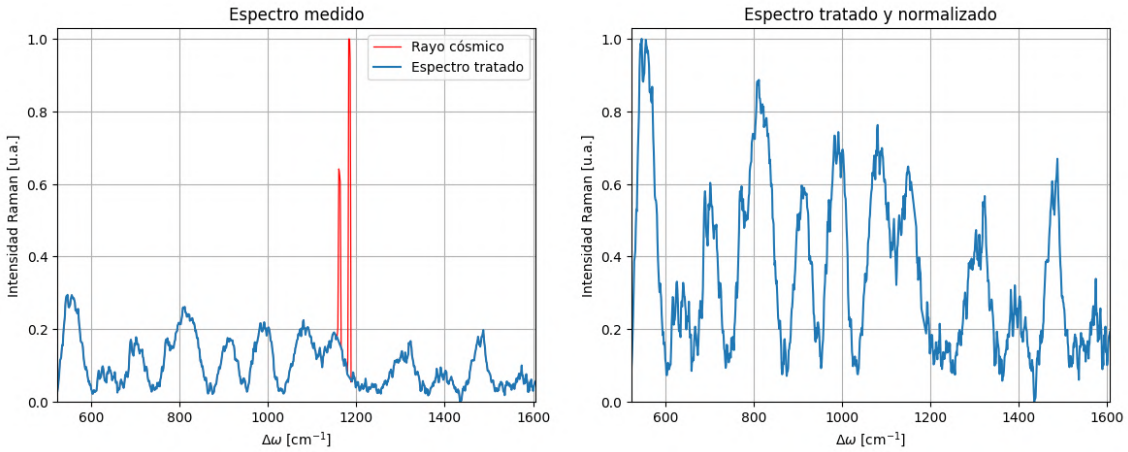

Figura 4.9. Ejemplo de la aplicación del algoritmo.

Como se puede comprobar, el resultado es excelente. No obstante y como se puede esperar, este procedimiento no elimina todos los spikes cuando se aplica a escala masiva en un dataset de este tamaño. Para agilizar la eliminación de spikes restantes se ha programado una aplicación de navegador utilizando la librería *dash* de Python que permita ajustar los parámetros del algoritmo y observar el resultado en tiempo real. Esta aplicación va a recibir y leer un archivo CSV con los datos de los espectros previamente corregidos para poder proceder a mejorar el resultado inspeccionando visualmente los resultados y operando directamente sobre ellos.

La interfaz de la aplicación tiene las siguientes funcionalidades:

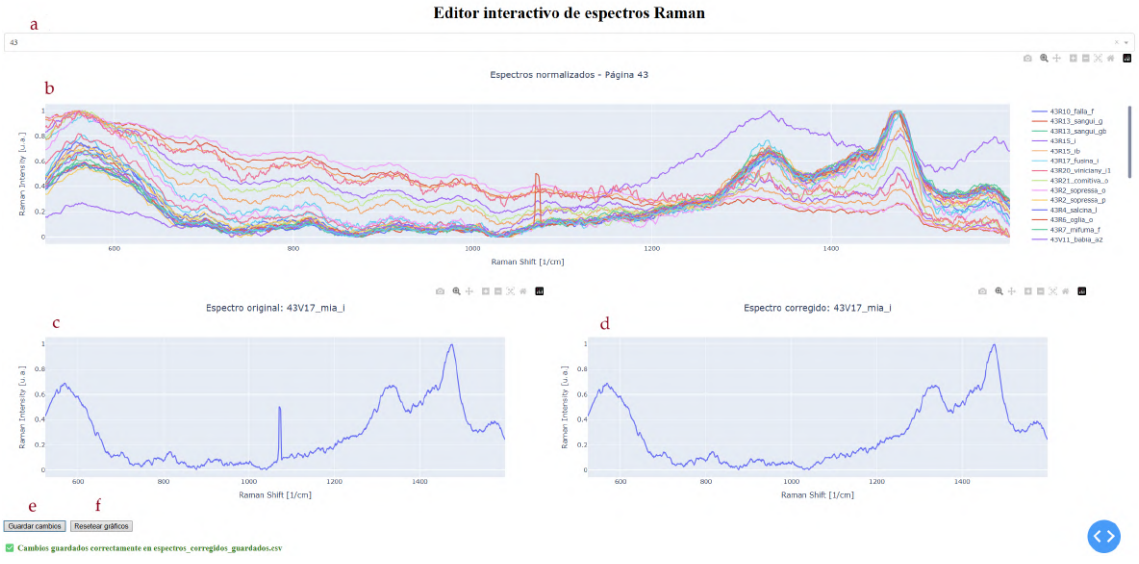

Figura 4.10. Interfaz de la aplicación Dash.

En esta interfaz:

a) Barra de búsqueda interactiva para seleccionar la página de interés.

b) Gráfica que muestra todos los espectros de esa página, con una leyenda interactiva para seleccionar qué espectros mostrar. Esta gráfica también es interactiva: permite seleccionar un espectro de interés haciendo click en él.

c) Gráfica que muestra el espectro seleccionado en **b**. Esta es la gráfica más importante: al hacer click sobre el pico que visualmente se identifica como un spike, se aplica automáticamente sobre ese spike el algoritmo aplicado en el punto anterior, modificando el espectro seleccionado.

d) Gráfica que muestra el resultado de la eliminación del spike en **c**.

e) Botón para guardar los cambios en el CSV.

f) Botón para resetear las gráficas.

Con estos pasos se obtienen datos limpios y listos para el análisis. A continuación, se evaluarán los parámetros de los distintos métodos de reducción de dimensionalidad para seleccionar los más adecuados, ya sea con el fin de generar un espacio fácilmente segmentable o de obtener componentes representativas de compuestos esperados. Posteriormente, se aplicará cada método y se comparará su rendimiento para decidir cuáles son útiles en el análisis.

Tras esta evaluación, se seleccionará el método más adecuado para continuar el trabajo. En primer lugar, se realizará el clustering con los algoritmos descritos y se calculará el espectro medio de cada cluster. Finalmente, se analizará la distribución espacial de los clusters en el manuscrito.

Capítulo 5. DISCUSIÓN

En esta sección se podrá ver el resultado de analizar con los métodos de reducción de dimensionalidad y clustering los espectros corregidos. Asimismo, se podrá comprobar más detenidamente el resultado del preprocesamiento de datos llevado a cabo.

5.1 Preprocesamiento

Para almacenar y nombrar los diferentes espectros Raman se ha seguido un estricto protocolo, teniéndose ejemplos como este:

25R2_fantonili_f

En este nombre, se tienen estas componentes:

- **25**: página del manuscrito en la que se toma la muestra. Este espectro se tomó en la página 25.
- **R2**: línea de la página en la que se toma la muestra. Se hace una distinción entre R (recto) y V (verso). En este caso, la muestra es de la segunda línea del recto de la página. Se puede ver la diferencia en este esquema:

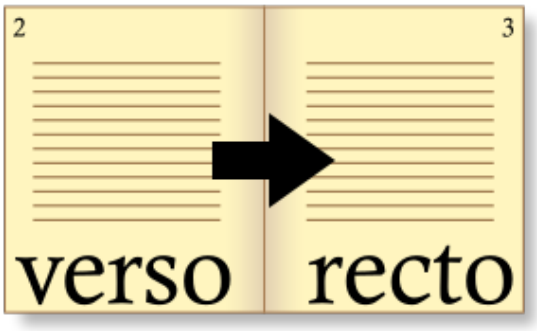

Figura 5.1. Recto y verso en un libro.
[32]

- **fantonili**: palabra en la que se toma la muestra.
- **f**: letra de la palabra de la cual es la muestra.

Como se mencionó previamente, hay que hacer una interpolación para que todos los valores del desplazamiento Raman sean los mismos y sean, así, comparables. Al acabar con eso, hay que aplicar el algoritmo automático de despiking (eliminación de spikes) y luego habría que limpiar los restantes manualmente con la aplicación.

Para el uso de la aplicación, es de especial interés obtener para cada espectro la página en la que se ha tomado. Ello facilitará tareas posteriores como mapeos de componentes y espectros medios de clusters. Para conseguir eso, se hará uso de una función y una lambda sobre el dataframe:

```python
# Se añade la columna "Pagina"
def extraer_pagina(nombre):
    match = re.match(r"(\d+)([RVrv])",nombre)   # Expresión regular para capturar número y letra

    # Si se encuentra un match, se formatea la página
    if match:
        numero = match.group(1)         # Número de página
        lado = match.group(2).upper()   # Letra (R o V) en mayúscula
        return f"{numero}{lado}"        # Formato "NúmeroLetra"
    return None

df_interp_normalized['Pagina'] = df_interp_normalized.index.to_series().apply(extraer_pagina)
```

Figura 5.2. Procedimiento para obtención de las páginas.

Aquí se pueden ver algunos de los resultados de utilizar la aplicación creada:

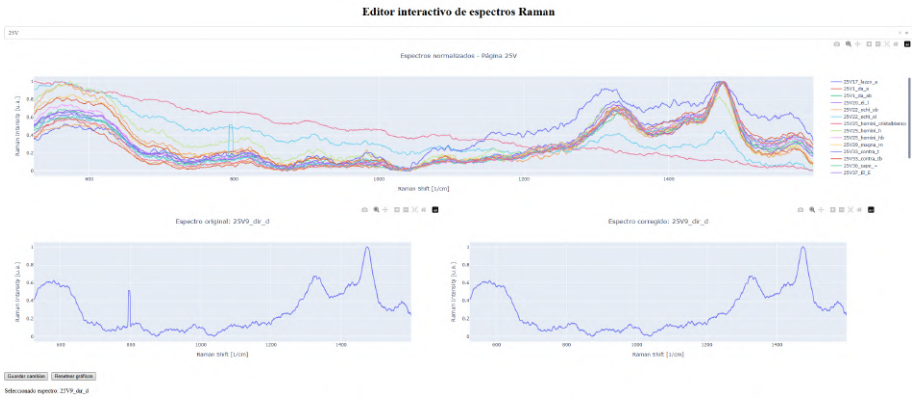

Figura 5.3. Despiking de un espectro de la página 25V.

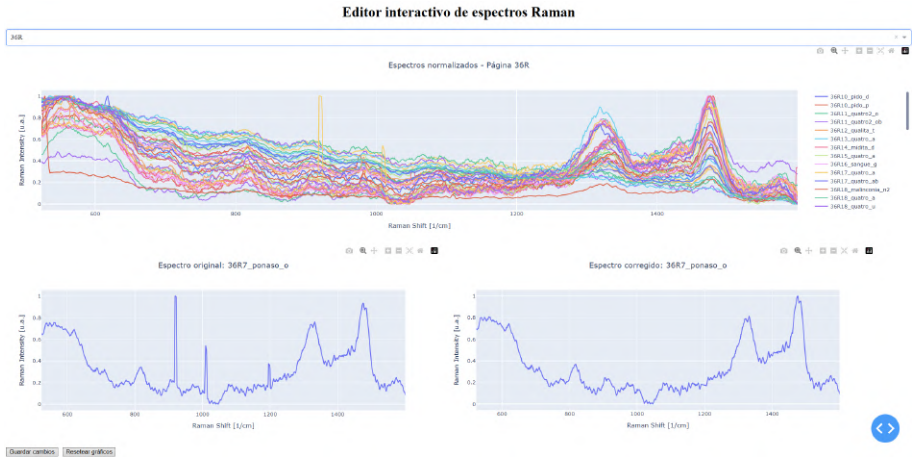

Figura 5.4. Despiking de un espectro de la página 36R.

Se puede ver que el resultado es muy bueno, lo que convierte esta aplicación en una poderosa herramienta para el tratamiento de espectros Raman.

El resultado de aplicar este procesamiento puede comprobarse en un caso particular, como el siguiente:

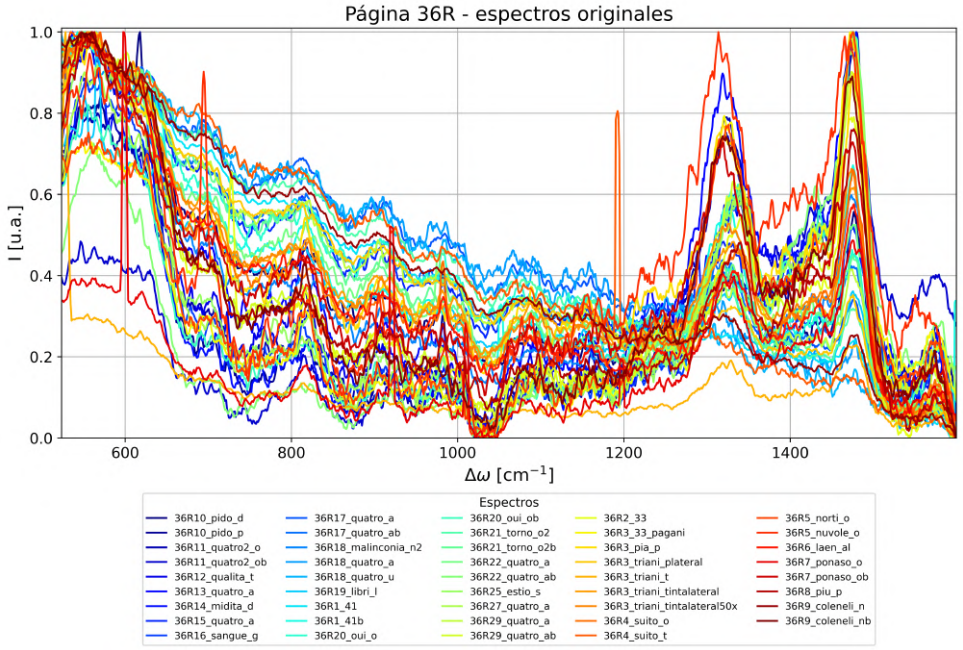

Figura 5.5. Espectros originales en la página 36R.

En esta imagen se puede ver una gran cantidad de spikes. Es algo que se puede esperar viendo el volumen de espectros que hay en esta página. Teniendo todo esto en cuenta, se aplicará el procedimiento en los dos pasos previamente mencionados, empezando por la localización y eliminación automáticas:

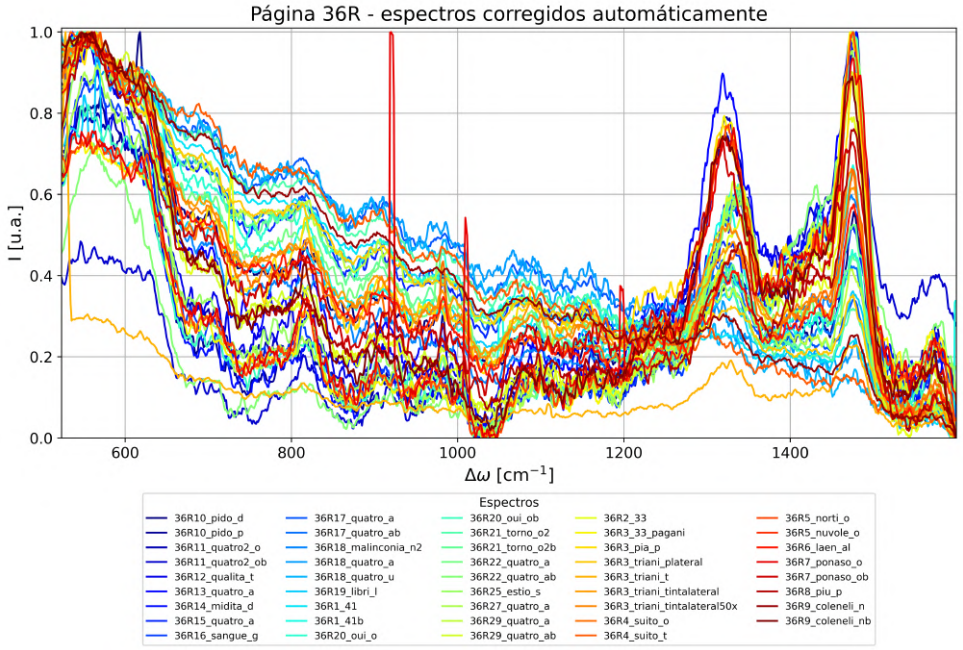

Figura 5.6. Espectros corregidos automáticamente en la página 36R.

Este primer paso ya ha eliminado la mayoría de los spikes presentes en la Figura 5.5. Hecho eso, se puede pasar a utilizar la aplicación desarrollada. Tras el tratamiento manual de los spikes restantes, el resultado final es este:

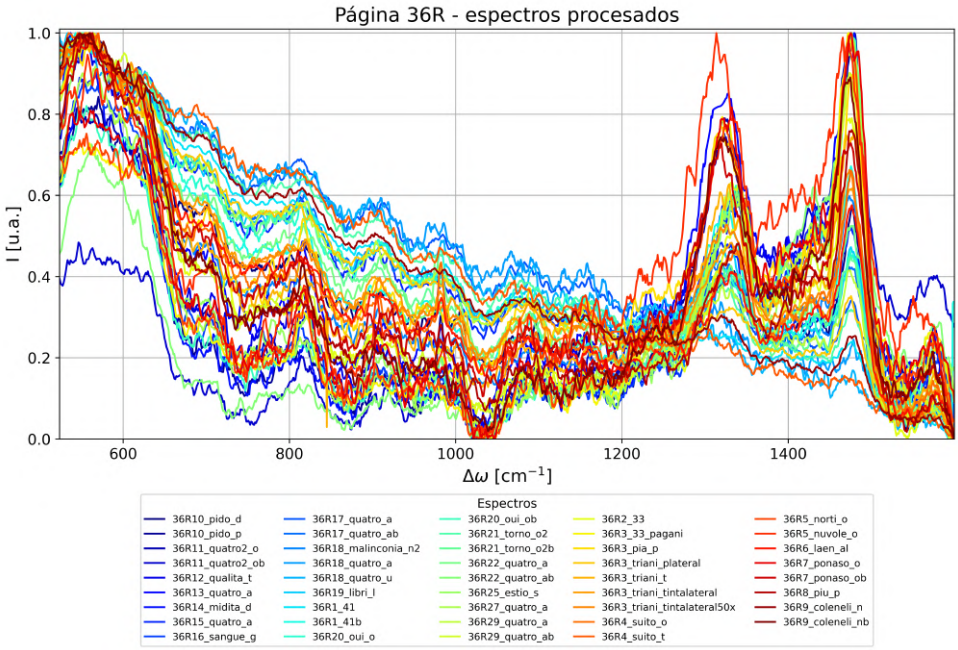

Figura 5.7. Espectros totalmente procesados en la página 36R.

Este proceso se aplica a todas las páginas y, hecho eso, se puede proceder ya con los métodos de reducción de dimensionalidad.

5.2 Resultados de t-SNE

Para aplicar este método de reducción de dimensionalidad se ha buscado un ajuste de diferentes parámetros de forma que se haga el espacio de embeddings lo más fácilmente segmentable posible.

Para determinar cuál es la mejor combinación de parámetros se ha llevado a cabo una inspección visual. El espacio más favorable para el análisis es el resultante de aplicar los siguientes parámetros:

```
tsne = TSNE(n_components = 2,          # Número de dimensiones de salida
            perplexity = 4,            # Perplejidad, controla la densidad de los puntos
            learning_rate = 1000,      # Tasa de aprendizaje
            n_iter = 20000,            # Número de iteraciones
            random_state = 0,          # Semilla para reproducibilidad
            init = 'pca')              # Inicialización con PCA para una mejor convergencia
```

Figura 5.8. Parámetros para el modelo t-SNE utilizado.

El espacio resultante es el siguiente:

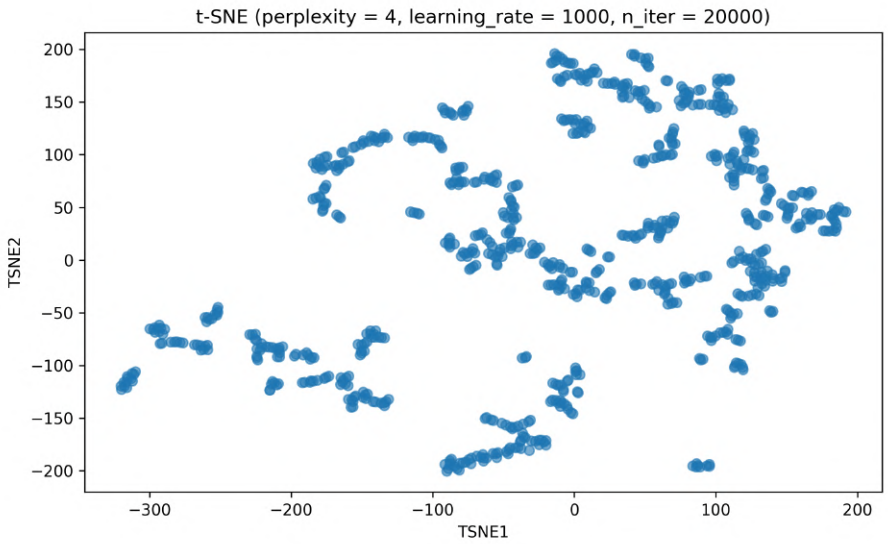

Figura 5.9. Espacio de dos dimensiones generado por t-SNE.

Pese a tardar más que otros métodos por el gran volumen de datos, t-SNE ha generado un espacio muy sencillo de clusterizar.

Ahora, para determinar los parámetros apropiados para el clustering en cada caso se revisarán las métricas descritas anteriormente, tal como se puede ver en la siguiente imagen para aplicar un clustering con K-Means:

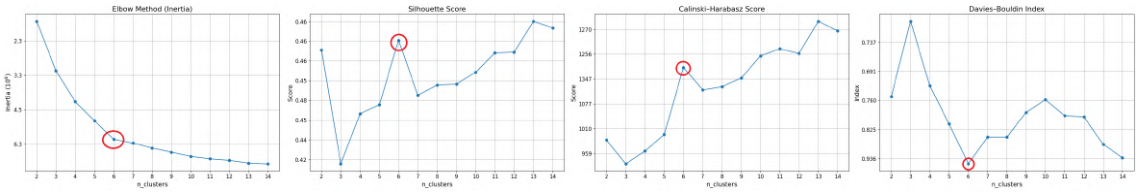

Figura 5.10. Métricas para definir el mejor número de clusters con K-Means.

Se puede ver de forma clara que el valor más apropiado para el algormitmo K-Means es $k = 6$ clusters.

Para los demás métodos de clustering seleccionados, se buscará la optimización de estos parámetros:

- **GMM**: se optimizará **n_components**.
- **Agglomerative Clustering - WARD/complete linkage**: se optimizará el parámetro **n_clusters**.
- **BIRCH**: de nuevo, se buscará el mejor **n_clusters**.
- **Spectral**: igual que en casos anteriores.
- **HDBSCAN**: se buscará el mejor **min_samples**.
- **OPTICS**: se tiene, de nuevo, **min_samples** como parámetro a ajustar.
- **AffinityPropagation**: en este caso, se intentará optimizar **preference**.

Hay muchos más parámetros que ajustar y que sin duda tienen un gran impacto en el resultado final, pero las implicaciones a nivel computacional de probar las distintas combinaciones de parámetros y el tiempo limitado de que se dispone para este trabajo hace inviable un estudio más riguroso en este sentido.

En este caso particular ha habido que descartar algunos métodos de clustering, puesto que generaban más de 100 clusters y eso, en el contexto de estos datos y lo que se sabe de ellos hasta ahora, es totalmente inviable porque las tintas ferrogálicas, por la naturaleza que tienen, no son tan diferentes unas de otras como para generar tantos grupos distintos. Probablemente entonces el algoritmo está granularizando el espacio en exceso. La segmentación generada por OPTICS no es tan abultada, si bien no es representativa en absoluto.

El resultado para los distintos métodos de clustering seleccionados se muestras en la siguiente figura:

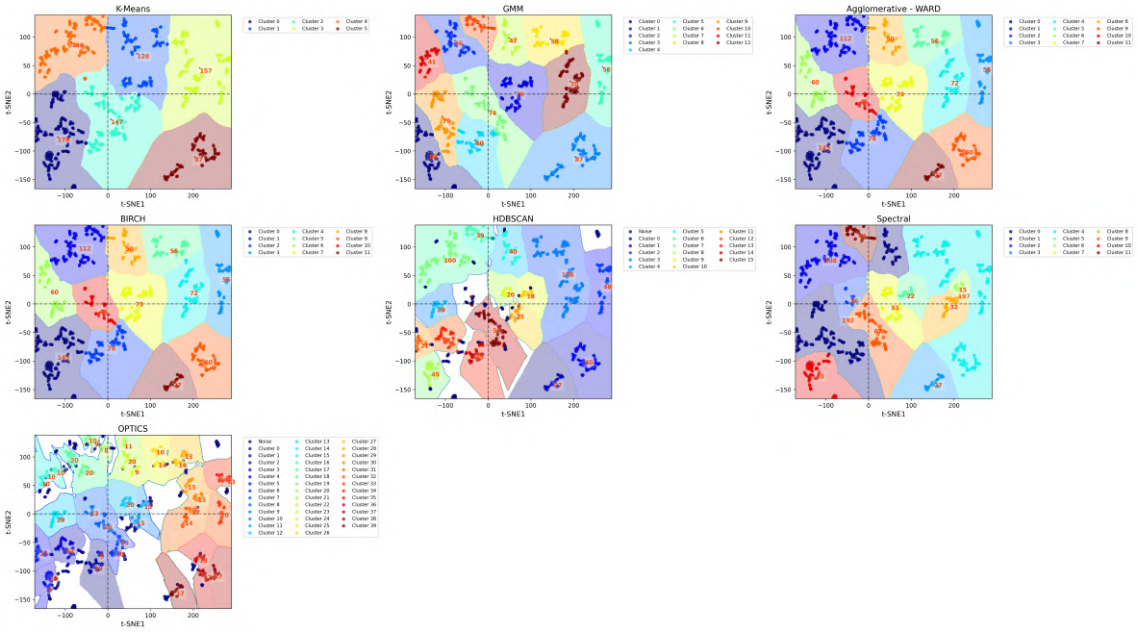

Figura 5.11. Clustering del espacio generado con t-SNE.

Se pueden extraer las siguientes conclusiones sobre estos métodos:

- Tanto K-Means como Spectral y BIRCH han conseguido una buena segmentación del espacio reducido. Las segmentaciones que hacen son claras y consistentes con las densidades de puntos.
- Agglomerative Clustering con WARD linkage tiene resultados similares a los de K-Means, mostrando diferencias en las fronteras entre clusters.
- GMM genera clusters más difusos con límites menos nítidos. No obstante, esto podría reflejar de forma más fiel la superposición entre clases.
- HDBSCAN muestra bastante ruido y varias regiones indefinidas.
- OPTICS muestra lo mismo que HDBSCAN pero con un espacio mucho más fragmentado, con una granularidad totalmente excesiva.

5.3 Resultados de UMAP

Al aplicar el método UMAP, con un random_state $= 42$, se obtiene un espacio de embeddings como este:

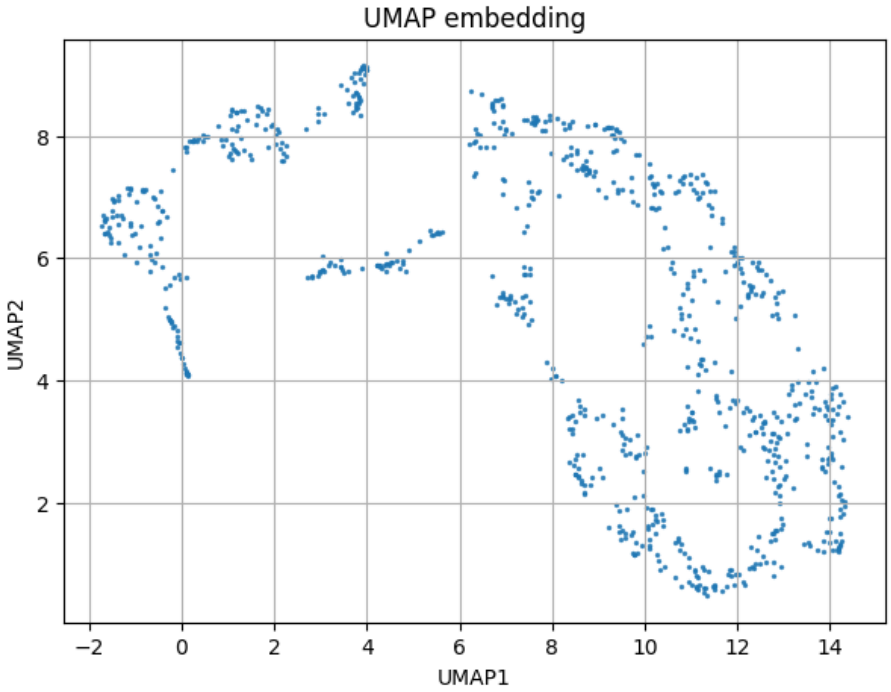

Figura 5.12. Espacio de embeddings generado por UMAP.

Este espacio puede ser segmentado por los métodos de clustering indicados anteriormente, obteniéndose (previo ajuste de hiperparámetros) un clustering como este:

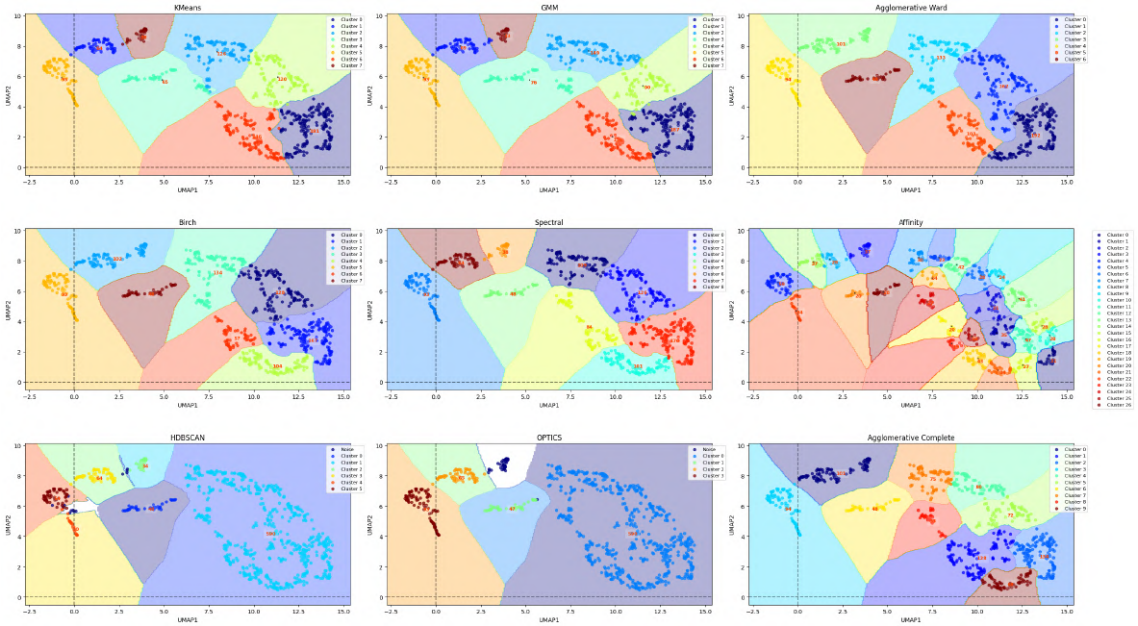

Figura 5.13. Clustering del espacio generado con UMAP.

En este caso:

- Los modelos K-Means, GMM, Spectral y BIRCH han segmentado regiones compactas y bien definidas. En este caso, las formas del nuevo espacio reducido favorecen el buen funcionamiento de GMM, cosa que no ocurría tanto con t-SNE.
- Los modelos de Agglomerative Clustering tienen un funcionamiento parecido al que mostraron para t-SNE pero, al igual que en el caso de BIRCH, muestran una separación más suave entre clusters.
- Affinity Propagation vuelve a crear más clusters de los necesarios.
- En los métodos de densidad (HDBSCAN y OPTICS) se forma un supercluster en el grupo que hay a la derecha del espacio reducido. Eso no tiene mucho sentido en este contexto, pues los espectros muestran una mayor variabilidad que la vista en estos casos.

5.4 ¿Métodos de embedding para este tipo de problemas?

Como se ha podido comprobar, tanto t-SNE como UMAP generan espacios reducidos que se pueden segmentar de forma relativamente sencilla. No obstante, el hecho de que no generan componentes cuya concentración pueda ser analizada a través de las páginas del manuscrito limita mucho su uso en este tipo de situaciones.

Por lo tanto y a priori, parece más útil en este tipo de problemas el uso de algoritmos como PCA, NMF o MCR-ALS, que sí que generan componentes que pueden ser catalogadas incluso como compuestos químicos. Esto facilita el seguimiento de las mismas por todas las páginas del manuscrito y puede dar lugar a un análisis más profundo.

Además, para un estudio como este, en el que se cuenta con especialistas de varias disciplinas para analizar los datos y conclusiones, estos métodos de reducción de dimensionalidad pueden dar un extra de explicabilidad que los métodos como UMAP o t-SNE no pueden ofrecer en ningún caso.

5.5 Resultados de PCA

Para hacer la reducción de dimensionalidad con PCA, lo primero será determinar el número de componentes a partir del cual se puede explicar una varianza significativa de los datos. Esto puede verse en las siguientes gráficas, donde se presenta la varianza explicada y la varianza cumulativa:

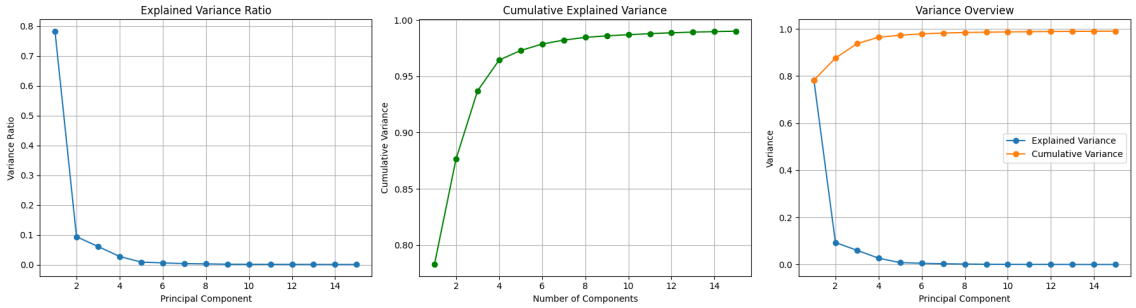

Figura 5.14. Evolución de varianza explicada por componentes principales de PCA.

Esto quiere decir que con 4 componentes se alcanza ya más del $95\,\%$ de varianza explicada. Para este análisis se tomarán 8 componentes principales coincidente con el $98\,\%$, ya que se esperan mas de cuatro componentes. Se pueden visualizar las relaciones entre ellas en este pairplot:

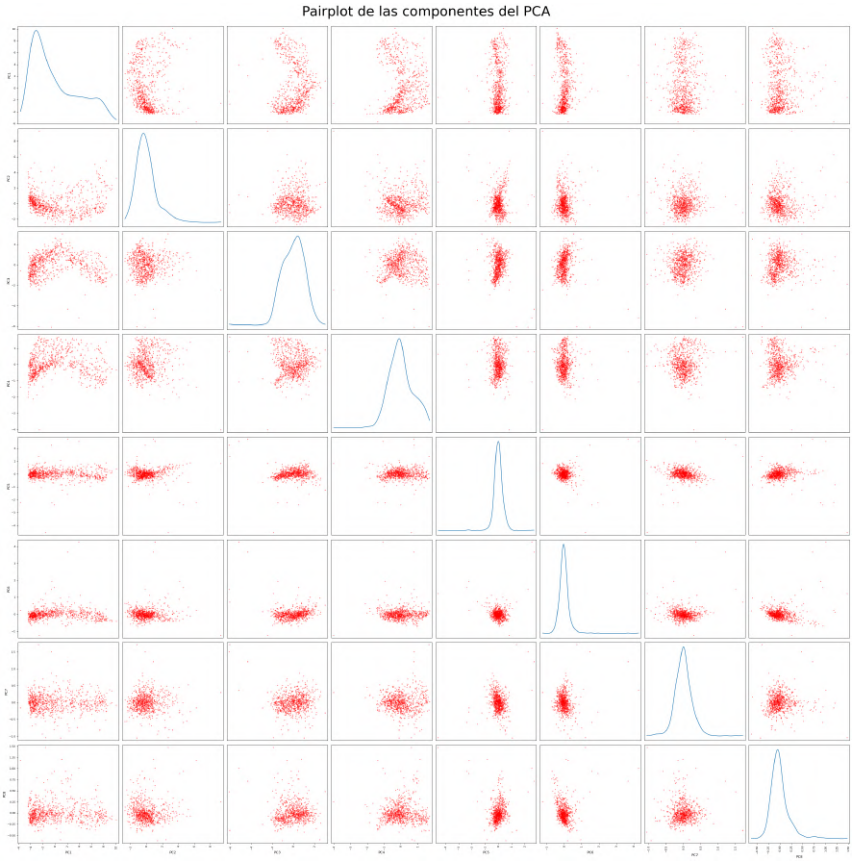

Figura 5.15. Pairplot de las componentes del PCA.

Varios pares de variables muestran curvaturas no lineales, lo cual podría sugerir que los datos no son lineales. Como el PCA sí es un método lineal, probablemente no sea el mejor método para estos datos. Estas son las componentes que genera este método de reducción de dimensionalidad:

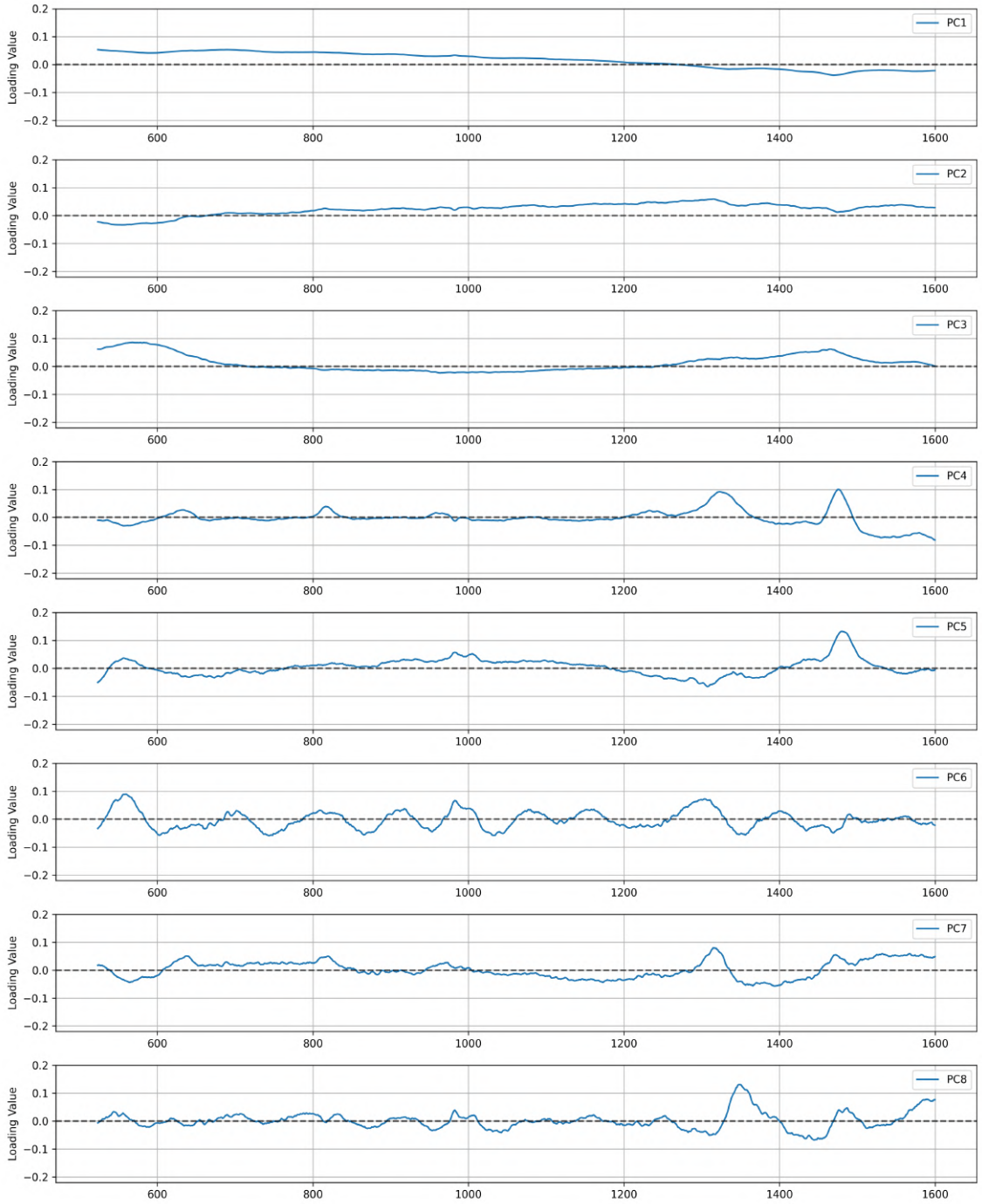

Figura 5.16. Componentes principales (PC) generadas.

Estas componentes no se corresponden con espectros asociados a compuestos químicos que puedan ser parte de las tintas ferrogálicas analizadas. Esto es debido principalmente

a que los componentes de PCA no tienen sentido químico-físico, ya que un espectro Raman real no debe tener valores negativos

Ahora, se hará el clustering sobre el espacio definido por las componentes principales determinadas. Para obtener una representación visual del mismo, se seleccionan PC1 y PC2, teniéndose los resultados siguientes:

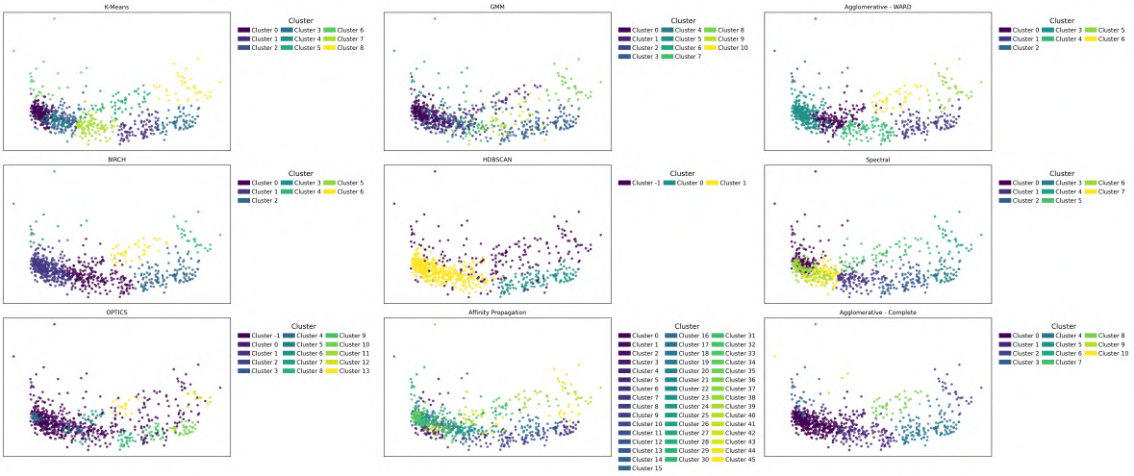

Figura 5.17. Clustering del espacio PCA.

Acerca de estas visualizaciones, se puede concluir lo siguiente:

- K-Means, BIRCH y Agglomerative Clustering - WARD/complete linkage generan clusters densos, contiguos y grandes. De estos cuatro, el que peor funciona es Agglomerative Clustering - complete linkage, que genera clusters con bordes menos claros.
- GMM genera mucho solapamiento entre clusters.
- HDBSCAN forma dos superclusters y luego el resto lo agrupa en ruido, lo cual es un muy mal resultado, en base a lo que se ve en los espectros.
- OPTICS genera también mucho solapamiento entre clusters.
- Affinity Propagation genera muchos microclusters de nuevo.
- Spectral separa la forma curva de forma nítida pero no se generan buenas transiciones entre clusters, generando problemas debido a la continuidad del espacio reducido.

5.6 Resultados de NMF

Determinar el número de componentes necesarias para utilizar NMF no resulta tan sencillo. El primer paso será analizar cómo va variando el error de Frobenius, para lo que se utiliza este código:

```
errors = []
ks = range(2,20)
spectral_df = df_interp[numeric_cols]
V = spectral_df.values
for k in ks:
    model = NMF(n_components = k,
                init = 'nndsvda',
                random_state = 42,
                max_iter = 10000)
    W = model.fit_transform(V)
    H = model.components_
    error = np.linalg.norm(V-W@H,'fro')
    errors.append(error)
```

Figura 5.18. Código utilizado para determinar el error de Frobenius.

Este código permitirá averiguar cómo varía dicho error para entre 2 y 19 componentes. En el caso del dataset original se utilizará un $max\_iter = 10000$ y para el renormalizado se hará con $max\_iter = 6000$.

Para ambos conjuntos de datos se tienen estos resultados:

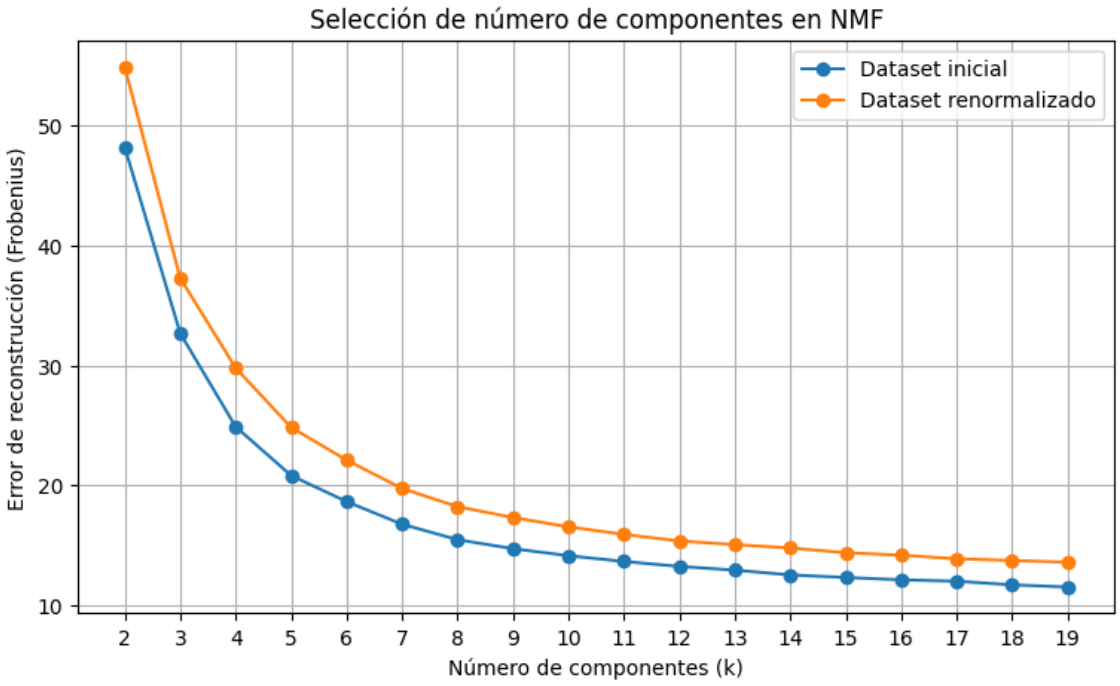

Figura 5.19. Evolución del error de Frobenius frente al número de componentes.

En esta gráfica puede verse que el error de Frobenius es ligeramente inferior en el caso del dataset inicial a aquel observado en el dataset renormalizado. Se puede también comprobar que a partir de 8 componentes el error decrece de forma muy lenta, conque se puede establecer este valor como el umbral para buscar el número óptimo de estos.

Para determinar el valor más apropiado, se procederá con una revisión de las componentes devueltas por el algoritmo, seleccionando la opción que devuelva componentes con mayor sentido físico-químico. De esta manera, se han hecho una serie de comprobaciones de forma sistemática:

- Un número bajo de componentes $(8 \leq N \leq 12)$ no genera unos componentes representativos de compuestos químicos de interés.
- Un número demasiado alto de esta variable $(N > 18)$ acaba por no generar tampoco componentes representativas, provocando una deformación de las componentes que sí podrían serlo.
- Para $max\_iter < 3000$, las componentes no convergen bien, independientemente del número de componentes seleccionado.
- Para $max\_iter > 7000$, no se observa una mejora significativa de las componentes generadas. Por tanto, no se observa una mejora visual ni analítica pero aumentar este parámetro sí influye significativamente en el tiempo de ejecución y rendimiento del programa.

Se pudo comprobar que desde el punto de vista químico, el resultado es óptimo para la siguiente combinación:

$$N = 15 \text{ componentes y } max\_iter = 5500 \text{ iteraciones}$$

Para esta combinación se han obtenido las siguientes componentes virtuales:

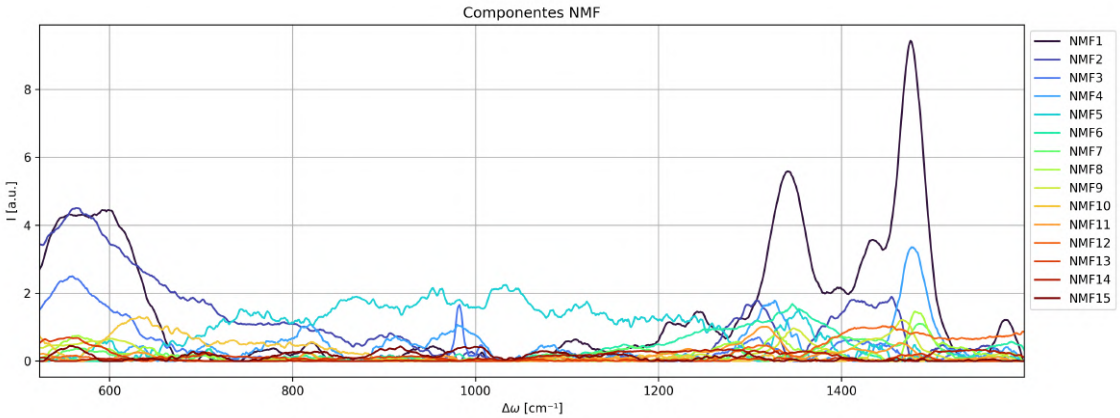

Figura 5.20. Componentes virtuales generadas por NMF.

La intensidad de cada componente depende del peso relativo con el que la señal aporta a la reconstrucción de los datos. Depende de la concentración relativa del compuesto y de la normalización ejecutada, por lo que no son concentraciones absolutas en ningún caso.

De la lista generada previamente se puede hacer una separación de los componentes que representan compuestos químicos reales y que, además, tiene sentido que aparezcan en este contexto:

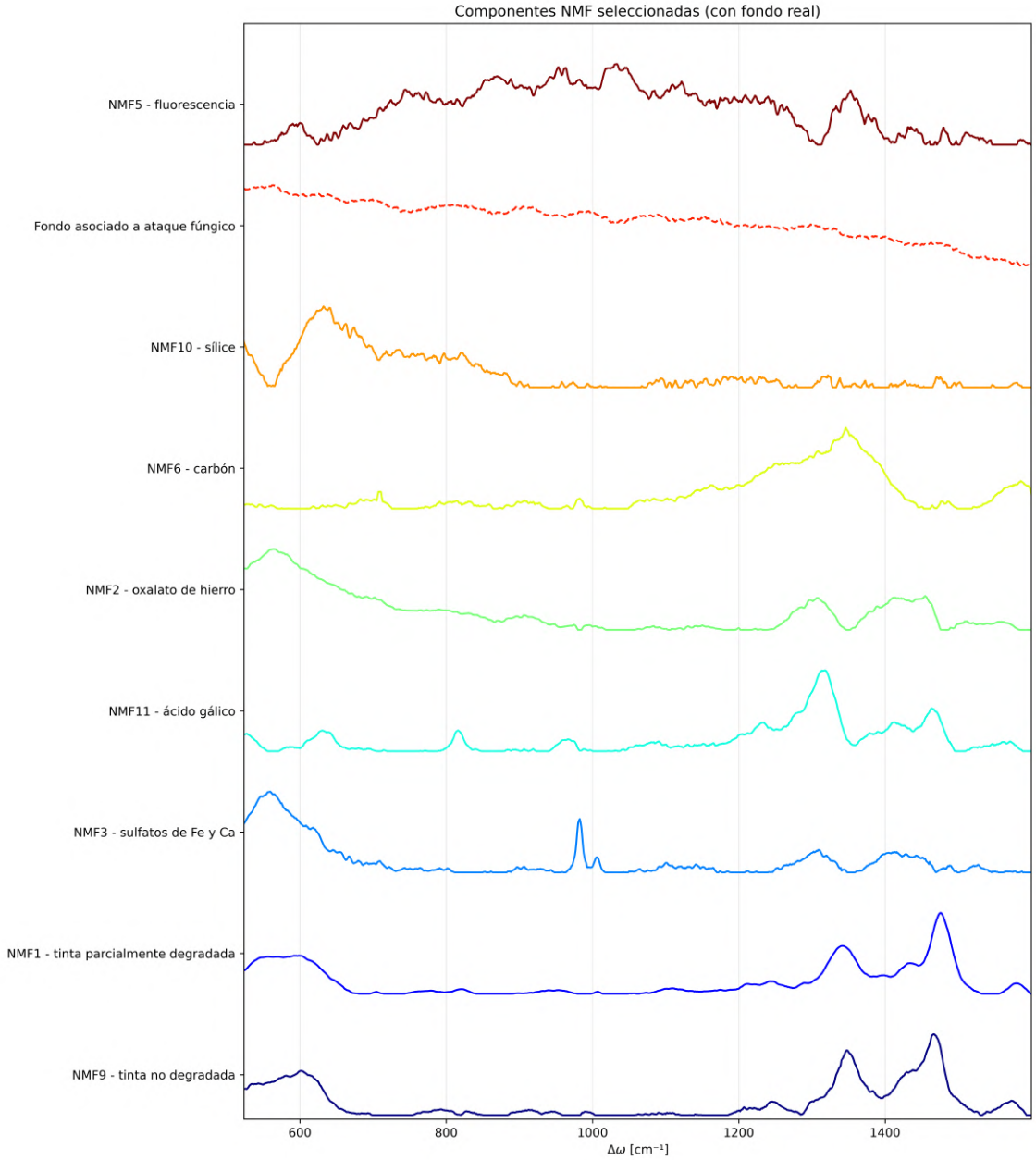

Figura 5.21. Compuestos químicos descubiertos con NMF y componente de fondo.

La componente de fondo está en línea discontinua porque no se detectó con NMF. De hecho, fue necesario depurar el dataset eliminando de él los espectros de este tipo para que NMF funcionara bien.

Análisis detallado de las componentes

Ahora, se puede analizar y ver por separado cada componente y cómo se distribuye por las distintas páginas:

- La **tinta no degradada** es aquella que muestra muy poca diferencia entre las alturas de las bandas que aparecen en la Figura 2.3 en los puntos $\Delta\omega \approx 1341\ cm^{-1}$ y $\Delta\omega \approx 1475\ cm^{-1}$, como se puede comprobar en esta imagen:

Figura 5.22. Componente NMF9.

Que aparezca esta componente indica que hay regiones del manuscrito donde se tiene una extraordinaria conservación de la tinta.

- La **tinta parcialmente degradada** aparece por la edad de la propia tinta y es la más común, como puede esperarse, en este manuscrito. Su espectro es el siguiente:

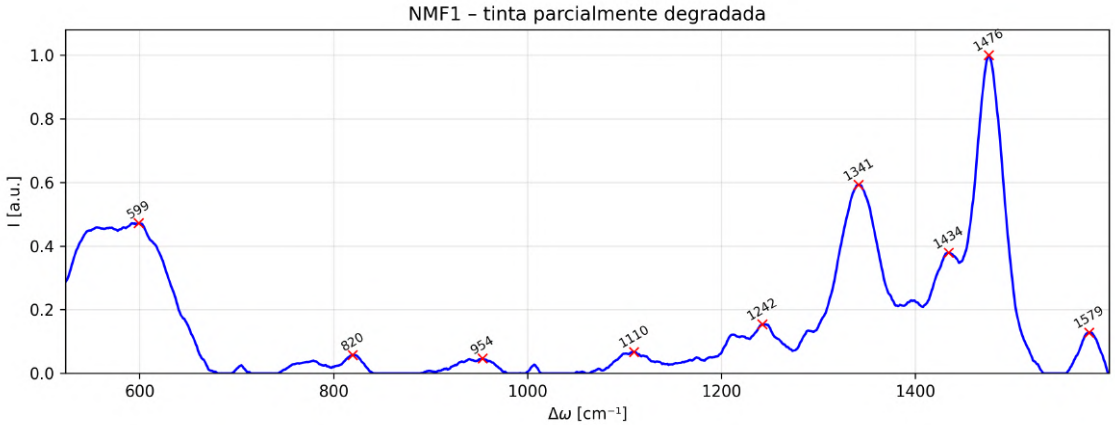

Figura 5.23. Componente NMF1.

En este caso se puede comprobar una mayor diferencia de altura entre las bandas anteriormente mencionadas. Esto es una observación llamativa porque podría indicar que se puede cuantificar cuánto y cómo se ha ido degradando la tinta con el tiempo gracias a la simple relación de altura entre estas bandas.

Las distribuciones por páginas de los dos componentes asociados a la tinta se pueden ver en estas gráficas:

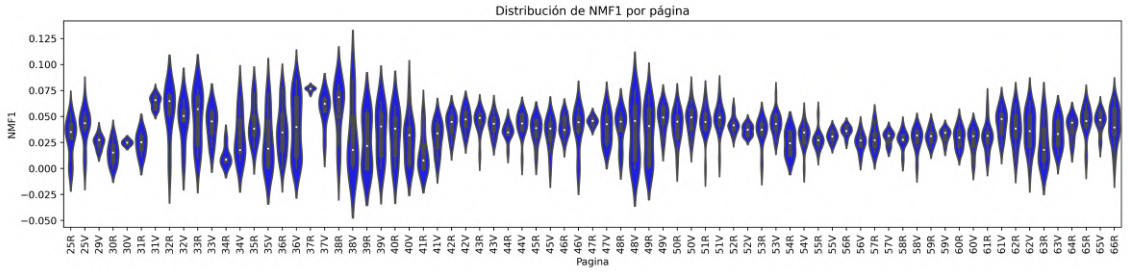

Figura 5.24. Gráfico de violín para la tinta parcialmente degradada.

La tinta parcialmente degradada presenta el siguiente rango: $-0{,}050 \leq$ NMF1 $\leq 0{,}125$ Esto implica una distribución muy homogénea por todo el manuscrito.

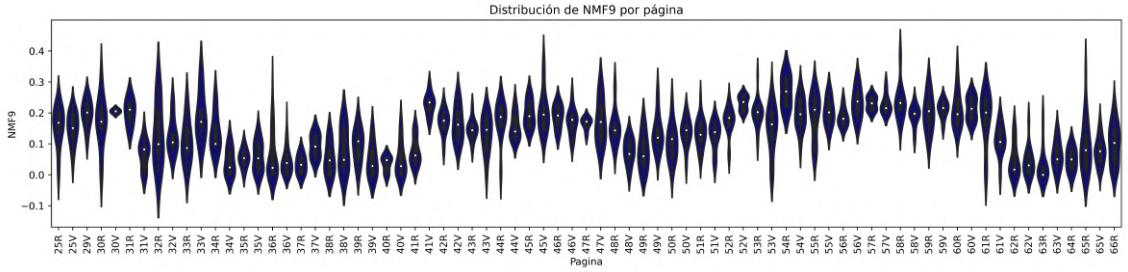

Figura 5.25. Gráfico de violín para la tinta no degradada.

Se puede comprobar que la tinta no degradada tiene una señal Raman más diversa, atendiendo al rango que presenta su gráfica de violín $-0{,}10 \leq$ NMF9 $\leq 0{,}45$.

La componente de tinta no degradada (NMF9) presenta una gran variabilidad entre páginas. Está especialmente presente en el principio del manuscrito (páginas 25 a 33), con violines anchos y medianas altas, y en el final (páginas 61 a 66), caracterizada por colas superiores muy largas. En la región central esta componente es más homogénea y débil.

Por contraparte, la tinta degradada tiene violines estrechos y medianas estables, lo que indica una distribución mucho más homogénea de este estado de la tinta por todo el manuscrito.

- Los **sulfatos de hierro (Fe) y calcio (Ca)** son muy importantes en cuanto a la conservación de los manuscritos.
  El sulfato de hierro aparece porque las tintas ferrogálicas se preparan con sales de hierro y extractos vegetales que contienen compuestos como el ácido gálico y los taninos. Este compuesto se caracteriza por una banda muy aguda en la región $980 \leq cm^{-1}\Delta\omega \leq 1000\ cm^{-1}$.
  El sulfato de calcio está presente en el soporte del manuscrito, en este caso papel. Este otro compuesto se caracteriza por una también muy aguda (pero menos que la de Fe, porque el espectro se toma sobre la tinta y no sobre el papel) en la región $1000 \leq cm^{-1}\Delta\omega \leq 1020\ cm^{-1}$. Esta es la componente:

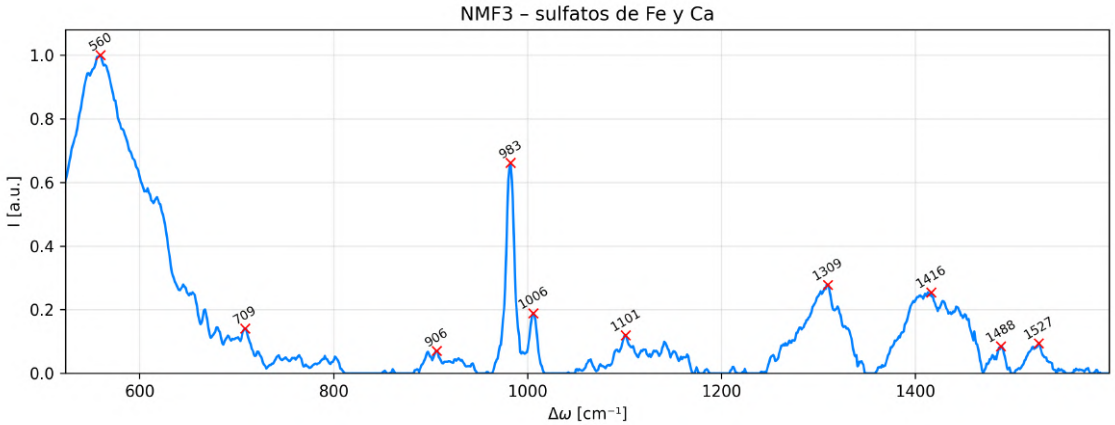

Figura 5.26. Componente NMF3.

El hecho de que estén estos compuestos químicos es importante porque indican daño en el papel: los sulfatos, cuando entran en contacto con la humedad, desestructuran la superficie del papel y el hierro favorece una serie de reacciones de reducción-oxidación (rédox) que dañan la celulosa. Ambos acidifican el soporte por liberación de H^+ al hidratarse, favoreciendo la hidrólisis de la celulosa, cosa que hace más frágil el sustrato
Es una componente muy interesante porque también aparecen en ella bandas asociadas al oxalato de hierro (del que se hablará en el siguiente punto), lo que implica que esta componente muestra los sulfatos y sus residuos. Así se distribuye esta componente:

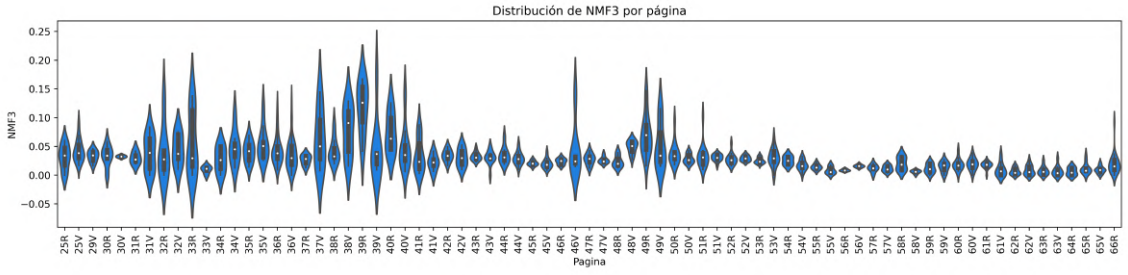

Figura 5.27. Gráfico de violín para los sulfatos de hierro y calcio.

Esta componente no es, por tanto muy frecuente, si bien aparece mucho entre las

páginas 31 y 33 y luego entre la 37 y la 41. En estas regiones, los violines se ensanchan mucho y tienen medianas altas. En el resto del manuscrito no tiene apenas incidencia. Este es un marcador mineral que indica una gran pureza del Fe usado en la tinta dichas zonas.

- El **ácido gálico** confirma el uso de extractos vegetales ricos en taninos para la preparación de esta tinta. Muestra bandas, de acuerdo con la Tabla 2.1, en $\Delta\omega \approx 816, 1430, 1478, 1580\ cm^{-1}$, todos ellos fácilmente detectables en esta gráfica:

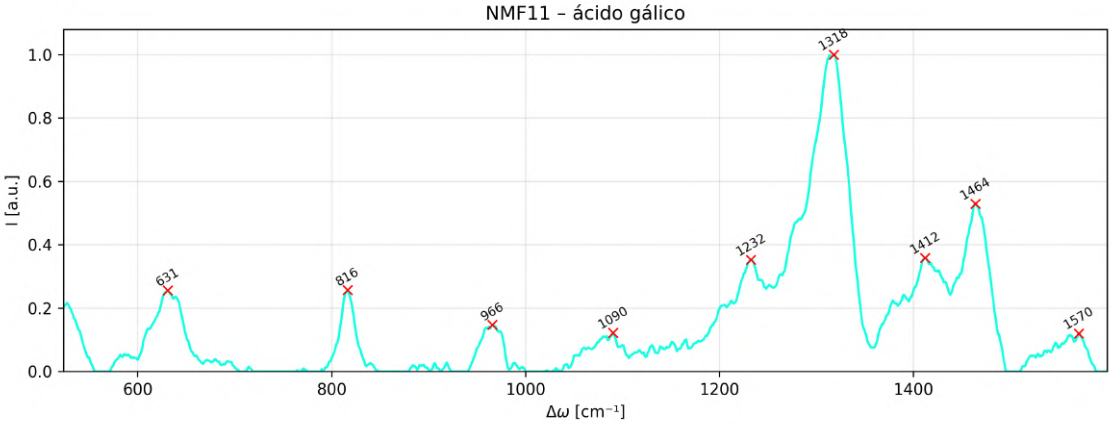

Figura 5.28. Componente NMF11.

Este compuesto acidifica el sustrato y acelera su degradación, conque también es una molécula de interés para la conservación de los manuscritos.

- El **oxalato de hierro** presenta picos característicos en las regiones $1300\ cm^{-1} \leq \Delta\omega \leq 1320\ cm^{-1}$ y $1400\ cm^{-1} \leq \Delta\omega \leq 1480\ cm^{-1}$, que permiten distinguirlo de sulfatos y ácido gálico. Estas bandas se pueden ver en esta imagen:

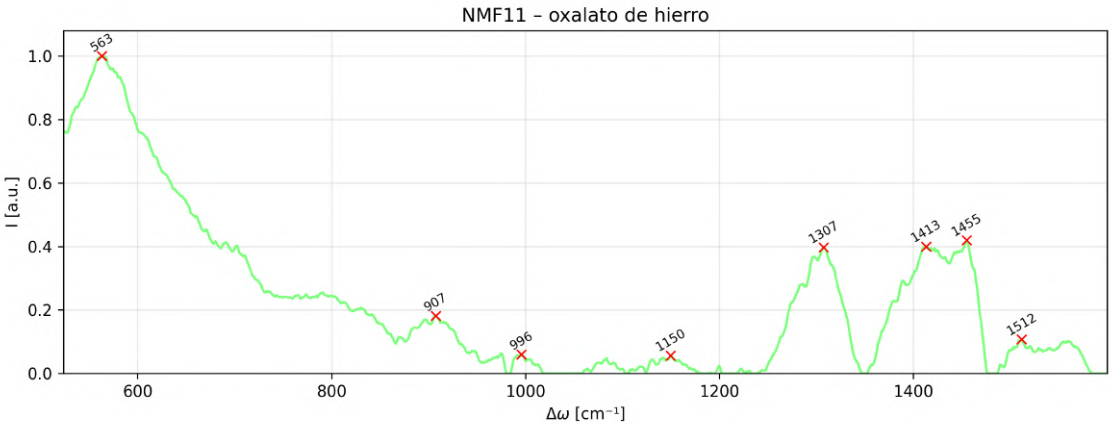

Figura 5.29. Componente NMF2.

El oxalato de hierro indica una catálisis del hierro y la fragmentación de moléculas orgánicas, ya sea en el soporte o en la tinta. Este compuesto puede, por tanto, ser utilizado como un indicador del grado de deterioro. Es una forma más estable del hierro

pero implica una pérdida de integridad a nivel químico del manuscrito.

Es, también, algo de interés el conocer cómo se han distribuido en este manuscrito estos dos compuestos:

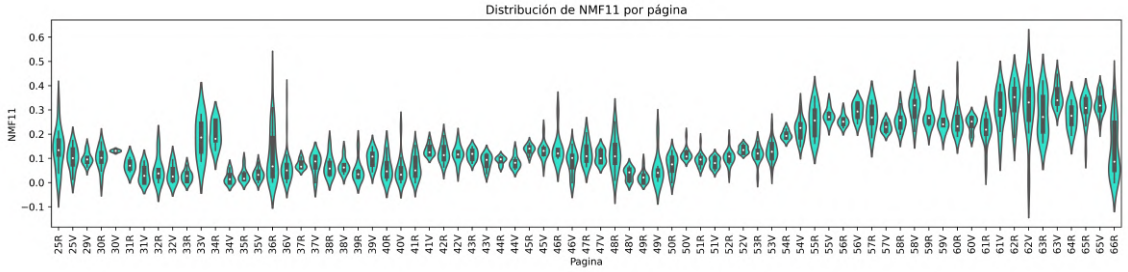

Figura 5.30. Gráfico de violín para el ácido gálico.

En este caso, la distribución tiene un rango muy amplio: $-0{,}1 \leq$NMF11$\leq 0{,}6$, con violines asimétricos. Esto indica una distribución muy irregular y dispersa.

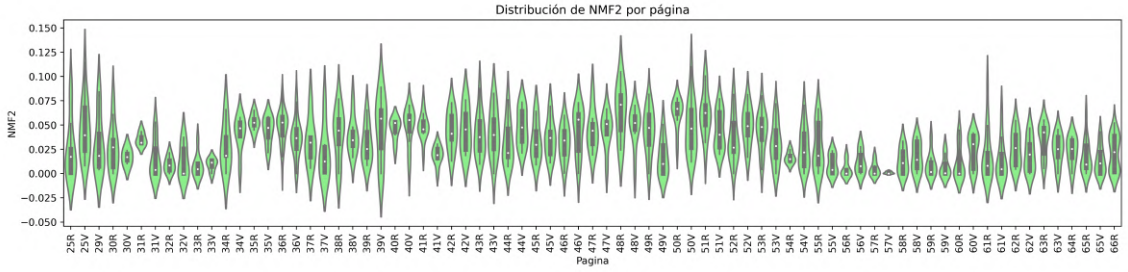

Figura 5.31. Gráfico de violín para el oxalato de hierro.

Esta distribución es mucho más homogénea en todo el manuscrito, con un rango de valores $-0{,}05 \leq$NMF2$\leq 0{,}15$. Las medianas muestran también poca variación.

En conjunto, ambas componentes reflejan diferentes facetas del deterioro de la tinta ferrogálica, analizadas también en un artículo de Kolar et al. [33]. Mientras que el oxalato de hierro aparece como un producto de degradación constante y cuya distribución es homogénea, el ácido gálico tiene una distribución más variable, tendiendo a concentrarse en las últimas páginas, que presentan violines amplios con largas colas superiores. Estas diferencias indican que los dos compuestos son consecuencia de diferentes procesos (los oxalatos aparecen asociados a microorganismos y el ácido gálico se debe a la hidrólisis de los taninos que forman parte de la tinta) y han de ser tenidos en cuenta para la correcta conservación del Patrimonio Histórico.

- El **carbón** es un compuesto fácilmente reconocible. Tiene una banda muy marcada en $\Delta\omega \approx 1350\ cm^{-1}$, que puede verse en esta ilustración:

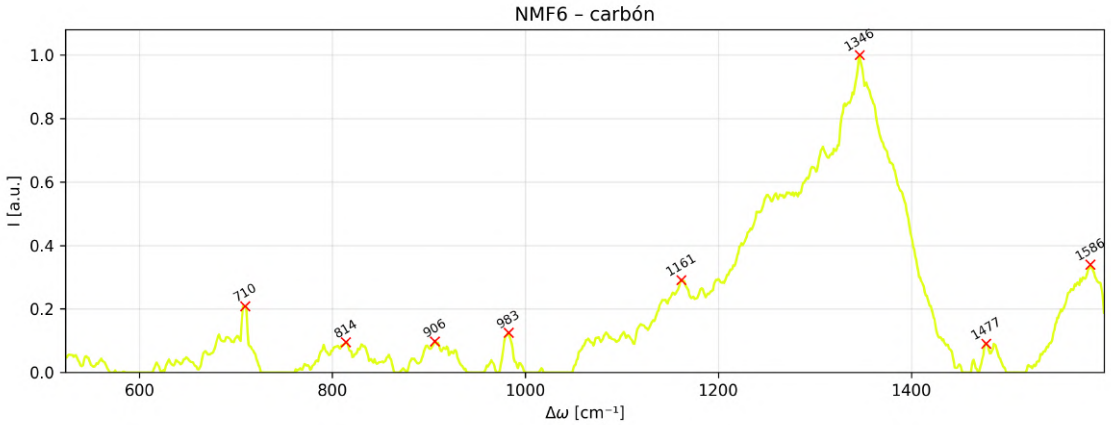

Figura 5.32. Componente NMF6.

Es un compuesto muy estable, por lo que no es un riesgo para la conservación, pero podría enmascarar otros compuestos de interés. Su presencia en este manuscrito podría deberse a una mezcla intencional o a una contaminación por, por ejemplo, escribir al lado de una vela.

- La **sílice** (dióxido de silicio - SiO_2) es un compuesto identificado por una fuerte banda en la región inicial del espectro, seguida de una caída continuada hasta el final. Es un compuesto que se halla en la arena.

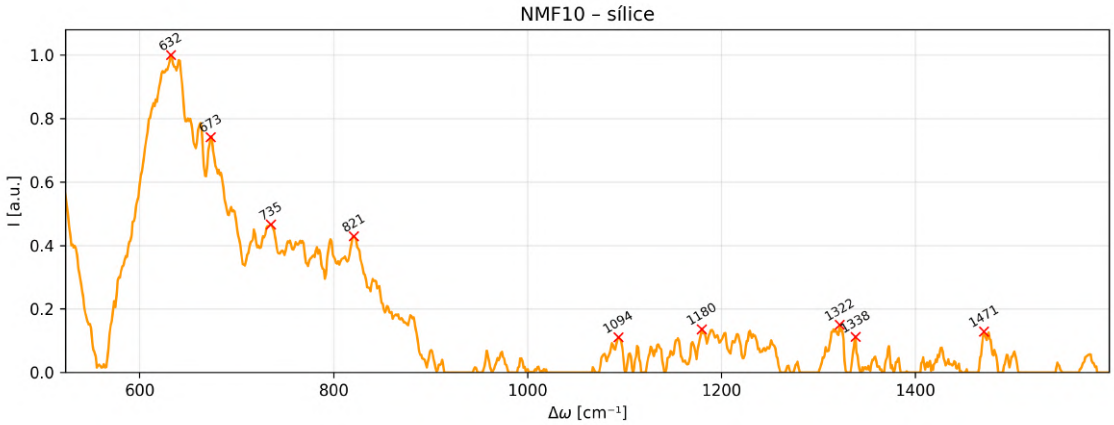

Figura 5.33. Componente NMF10

La presencia de estas moléculas en el manuscrito podrían indicar que se ha utilizado arena como agente secante para la tinta recién escrita. Esto era una práctica común en la época de la que data esta obra.

- Se ha observado una **colonización por microorganismos**, que deja un cierto fondo en algunos espectros. Estos pueden inducir una biodegradación adicional (oxidación, biomineralización...) en los taninos empleados al hacer la tinta. Estos efectos pueden dar lugar a la aparición de compuestos como el oxalato de hierro, que ha sido ya comentado previamente. Se puede ver en espectros como este:

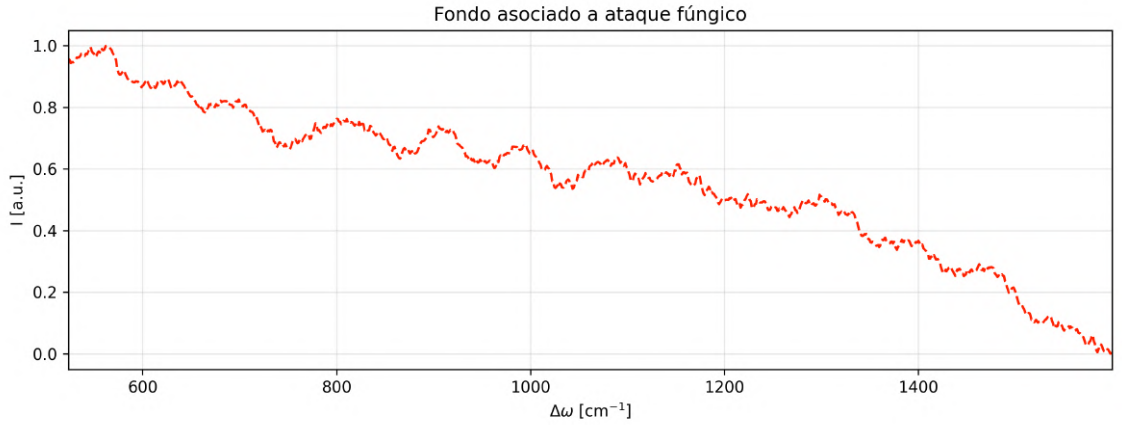

Figura 5.34. Espectro Raman del fondo por colonización de microorganismos.

Estos espectros pueden ser vistos en las siguientes páginas:

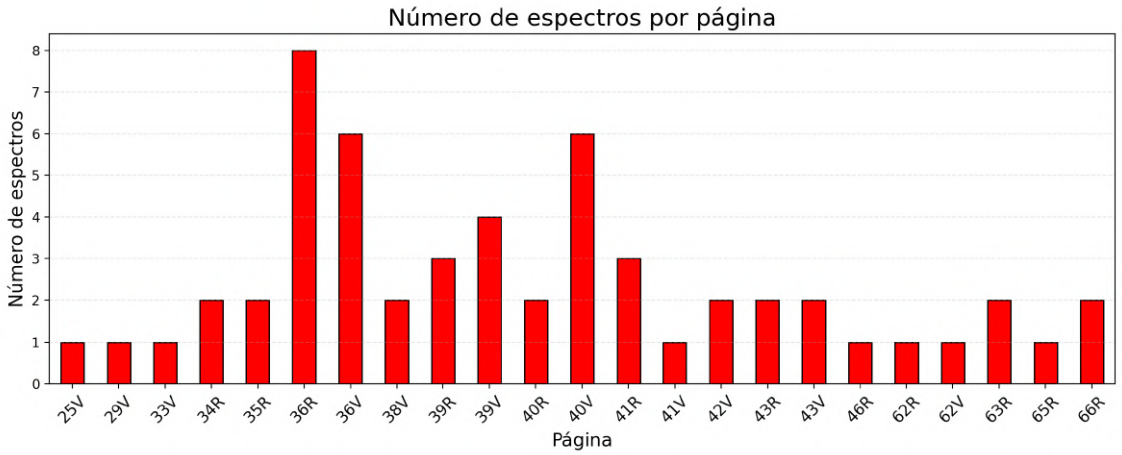

Figura 5.35. Distribución de los espectros de fondo.

Según esta nueva visualización, puede verse que estos espectros son especialmente frecuentes en páginas donde se observó una gran cantidad de sulfatos. Esto es algo que tiene mucho sentido pues, como se ha podido comprobar en artículos como el de del Castillo Rueda et al. [34], el hierro juega un papel fundamental en la aparición, supervivencia y virulencia de los microorganismos.

- Por último, se ha detectado **fluorescencia**, caracterizada por la ausencia de bandas definidas y una señal oscilante en todo el espectro. Este comportamiento se asocia a la presencia de productos de degradación orgánica y posibles metabolitos fúngicos. La fluorescencia constituye un marcador indirecto del grado de deterioro y de la complejidad química de la tinta y el soporte.

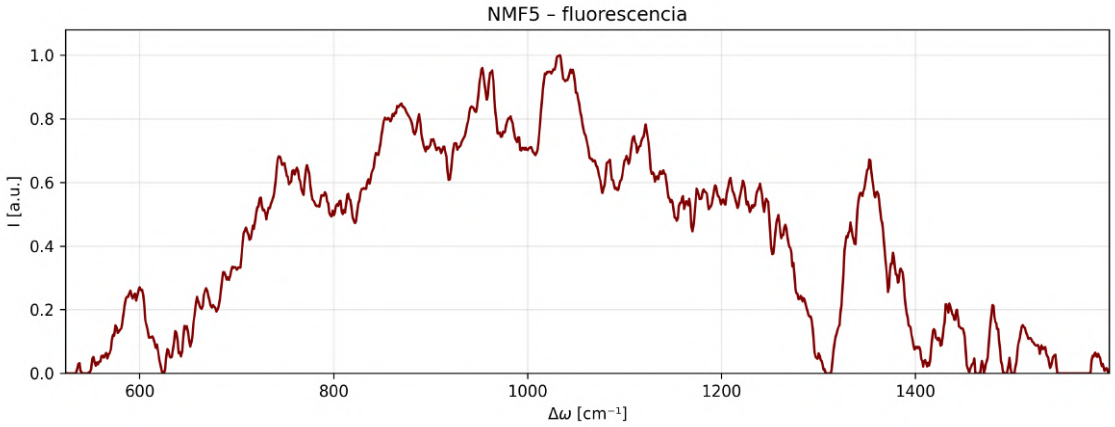

Figura 5.36. Componente NMF5.

La distribución por páginas de esta fluorescencia es la siguiente:

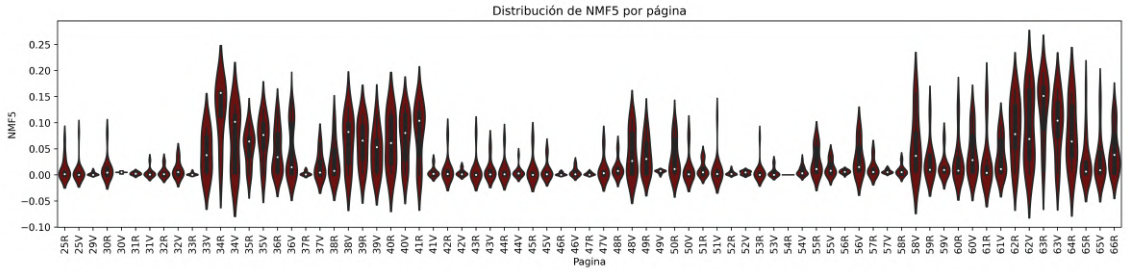

Figura 5.37. Gráfico de violín para la fluorescencia.

Como puede verse, aparece también en la región donde hay mucho hierro. Esto se debe al hecho de que los compuestos de hierrro son muy absorbentes en el visible, lo que hace que absorban parte de la luz láser y generen procesos de reemisión de la misma. Esto, en última instancia, provoca la aparición del fondo que se ve en la Figura 5.36. Este es un efecto conocido y bien documentado en artículos como el publicado en 2009 por Hanesch et al. [35].

Todas las componentes restantes que no se han mencionado son también fluorescencia, pero este era el ejemplo en el que mejor se veía.

Clustering

Para continuar el análisis, es imperativo obtener el análisis NMF completo, dividido en dos matrices. Para esto:

```
scores_nmf = nmf.fit_transform(df_interp.drop(columns = 'Pagina'))      # shape: (n_spectra,N_comp)
loadings_nmf = nmf.components_                                          # shape: (N_comp,n_features)
```

Figura 5.38. Código usado para el análisis NMF.

Se tienen dos matrices: las componentes (loadings) y las puntuaciones de cada uno (scores). Para hacer el clustering habrá que utilizar el dataset de puntuaciones. Sabiendo esto, se ha aplicado también en este caso el procedimiento visto en la Figura 5.10 para todos los métodos de clustering descritos. Los **mejores resultados** se han obtenido para los métodos **GMM** y **AgglomerativeClustering - WARD linkage**. Los demás métodos se han descartado por presentar clusters repetitivos/inconsistentes o una excesiva granularidad, siendo útiles solo en caso de querer detectar espectros especiales muy concretos. Con todo esto:

- **Gaussian Mixture Model (GMM)**: la aplicación de este método da lugar a 7 clusters diferentes (con uno muy minoritario), cuyos espectros medios son los siguientes:

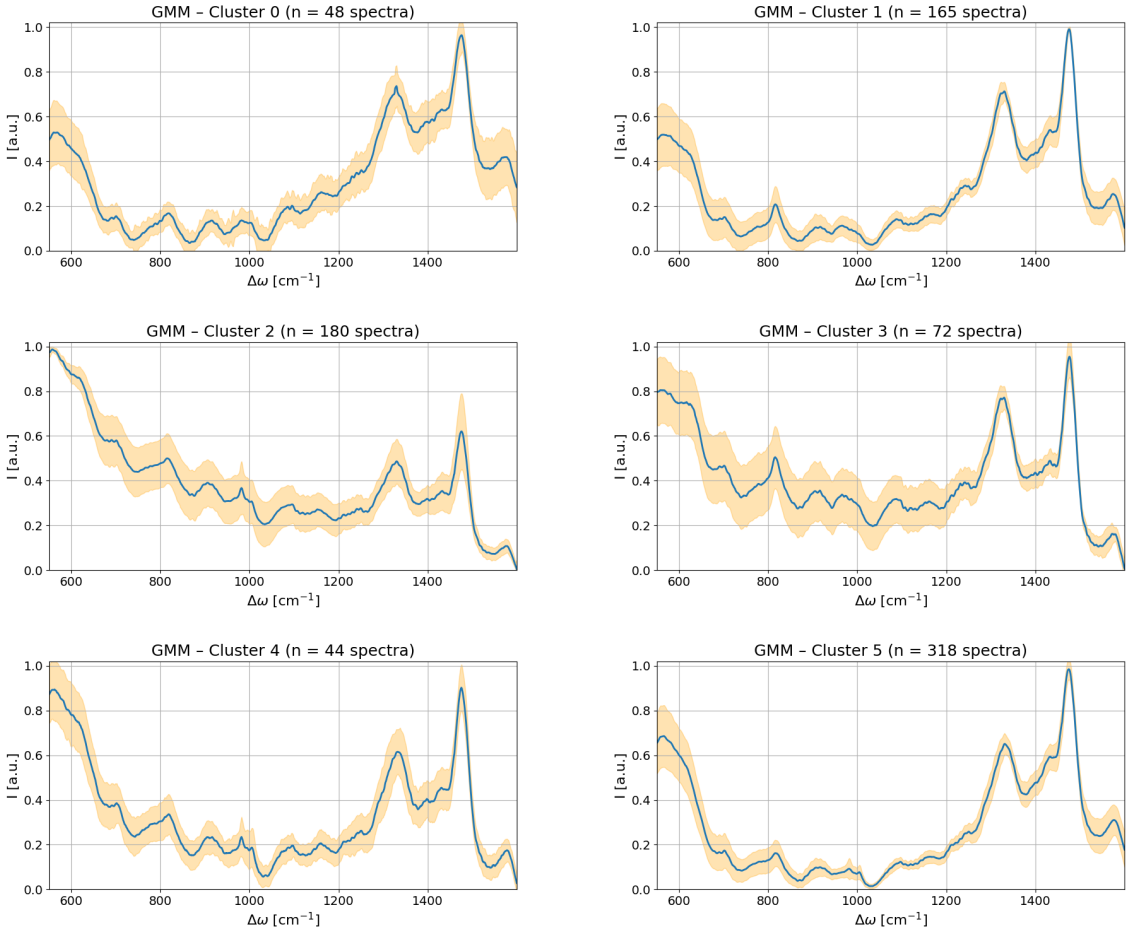

Figura 5.39. Espectros medios de los clusters generados por GMM.

- El **Cluster 0** corresponde a una tinta parcialmente degradada. No presenta muchas características definitorias más allá de las bandas de $\Delta\omega \approx 1350\ cm^{-1}$ y $\Delta\omega \approx 1450\ cm^{-1}$. Se tiene esta relación de altura entre ambas bandas:

$$\frac{\Delta\omega_{1350}}{\Delta\omega_{1450}} \approx 0{,}78 \pm 0{,}15$$

  Atendiendo a la desviación estándar en el último tercio del espectro medio, se puede ver que hay bastante ruido en este cluster en dicha región.
- El **Cluster 1** es un espectro de tinta parcialmente degradada. La relación de altura entre las dos bandas antes mencionadas es muy parecida, como se puede comprobar:

$$\frac{\Delta\omega_{1350}}{\Delta\omega_{1450}} \approx 0{,}72 \pm 0{,}05$$

  Este cluster presenta una menor dispersión que el anterior. No obstante, en este caso se puede ver la banda de $\Delta\omega \approx 816\ cm^{-1}$ de forma bastante clara, cosa que indica presencia de ácido gálico.
- El **Cluster 2** muestra una tinta menos degradada que los anteriores, pues se ve reducida la relación entre las bandas $\Delta\omega \approx 1350\ cm^{-1}$ y $\Delta\omega \approx 1450\ cm^{-1}$:

$$\frac{\Delta\omega_{1350}}{\Delta\omega_{1450}} \approx 0{,}82 \pm 0{,}12$$

  En este caso se ve de forma clara también la banda de $\Delta\omega \approx 950\ cm^{-1}$, lo que indica una notable presencia de sulfato de hierro.
- El **Cluster 3** tiene una relación entre las bandas igual a la del anterior, lo que indica una buena conservación de la tinta:

$$\frac{\Delta\omega_{1350}}{\Delta\omega_{1450}} \approx 0{,}82 \pm 0{,}09$$

  Este cluster también tiene una gran prominencia en la banda de $\Delta\omega \approx 816\ cm^{-1}$. Si bien la dispersión en la segunda mitad es escasa, en la primera mitad sí hay una notable desviación estándar.
- El **Cluster 4** presenta una relación entre las bandas muy cercana a la del Cluster 1, cosa que indica un estado avanzado de degradación de la tinta:

$$\frac{\Delta\omega_{1350}}{\Delta\omega_{1450}} \approx 0{,}71 \pm 0{,}13$$

  De nuevo, en este cluster se aprecia bien la presencia de hierro gracias a la banda de $\Delta\omega \approx 950\ cm^{-1}$. La desviación estándar de este cluster es muy baja.
- El **Cluster 5** es el más común de todos y presenta un espectro genérico de tinta ferrogálica parcialmente degradada. Se ven las bandas asociadas al ácido gálico y al hierro pero son poco prominentes. Por su parte, la relación entre las bandas de 1350 y 1450 es la siguiente:

$$\frac{\Delta\omega_{1350}}{\Delta\omega_{1450}} \approx 0{,}66 \pm 0{,}08$$

  Se puede concluir que las zonas que abarca este cluster son las más degradadas de todo el manuscrito y, además, todos los espectros están en un estado de similar degradación, dada la baja desviación estándar presente en el mismo.

- El **Cluster 6** tiene solo dos espectros y son de fluorescencia, como se puede ver:

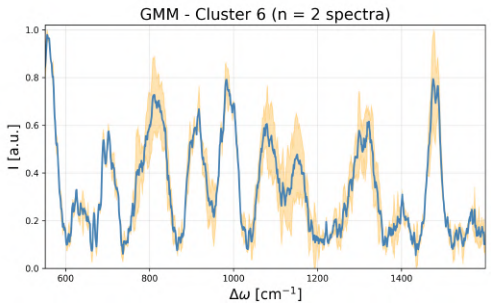

Figura 5.40. Espectro medio del Cluster 6 generado por GMM.

También se puede analizar cómo se distribuyen estos clusters a través de las páginas:

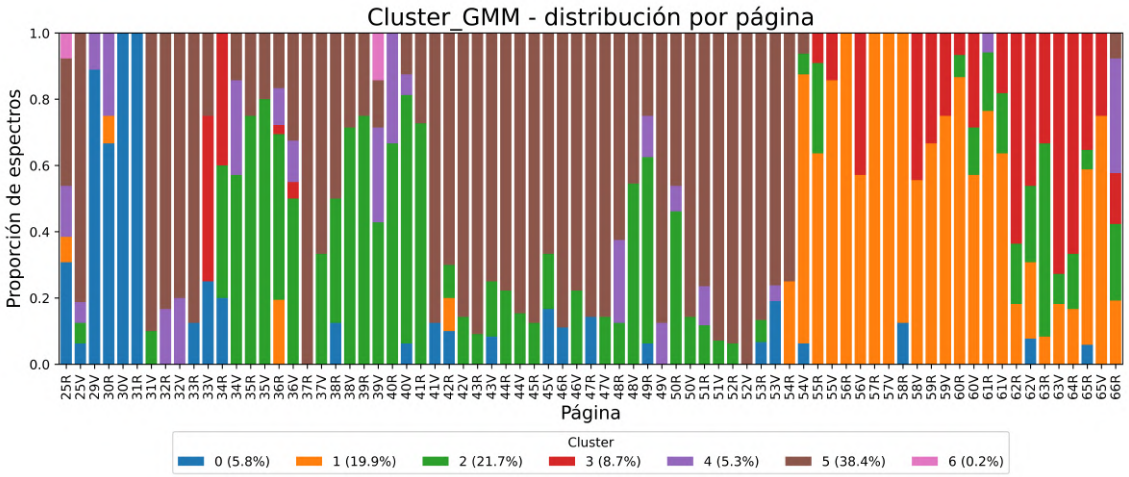

Figura 5.41. Distribución por páginas de los clusters generados por GMM.

Se puede comprobar que el cluster más frecuente es, con diferencia, el Cluster 5. Este domina de forma muy amplia desde el principio del manuscrito hasta la página 54R, desde donde deja de aparecer abruptamente.

En su lugar, al final del manuscrito el que domina es el Cluster 1. Esto indica una mayor presencia de ácido gálico en estas partes de la obra, lo cual es absolutamente consistente con lo visto en la Figura 5.30. El Cluster 3, que muestra también ácido gálico, aparece en algunas páginas al inicio del manuscrito pero más en concreto al final. Este es el cluster donde se observa una mejor conservación de la tinta.

El Cluster 2, el segundo más común (y con buen estado de conservación de la tinta también), aparece en gran medida desde la página 34R hasta la página 41R y posteriormente en páginas sueltas como la 48V y la 49R, así como hacia el final. Este cluster mostraba presesncia de hierro. Este mismo comportamiento se puede comprobar en la Figura 5.31. El otro cluster que contiene sulfato de hierro, el Cluster 4, aparece esporádicamente en el principio y el centro del manuscrito.

Hay una página especialmente interesante, que es la 36R. En ella, se han detectado diferencias entre las tintas en líneas pares o impares, como se puede ver:

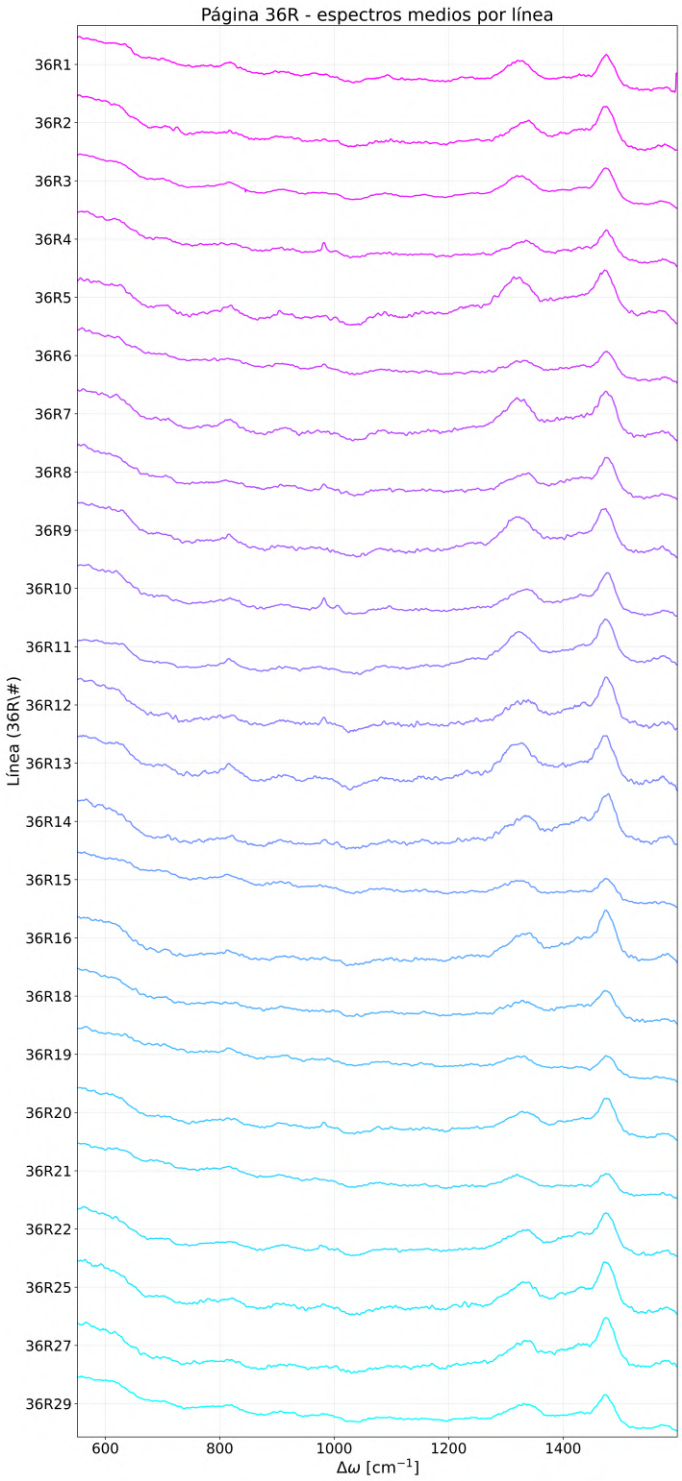

Figura 5.42. Espectros medios por línea en la página 36R.

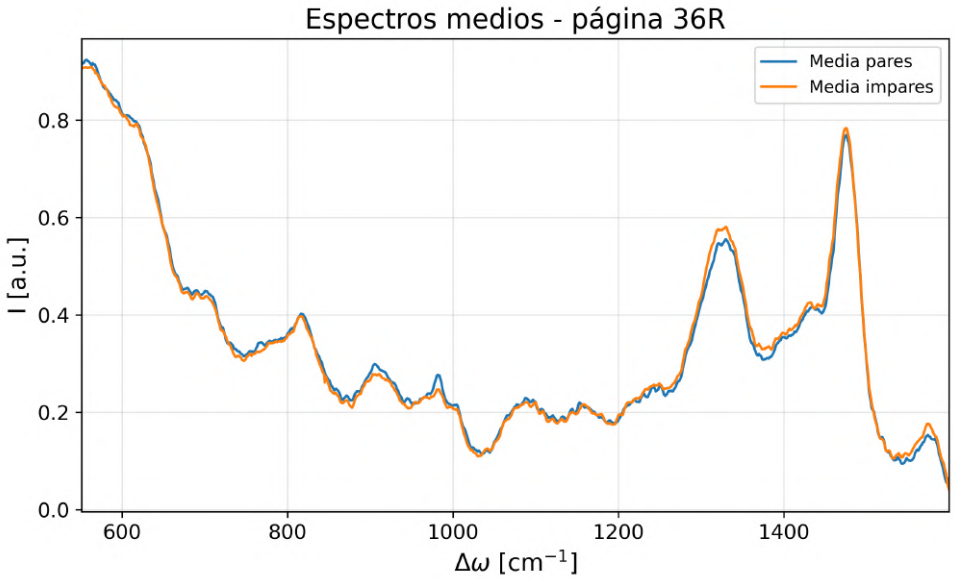

Figura 5.43. Espectros medios de líneas pares e impares en la página 36R.

Obteniendo el espectro medio de las líneas pares y las impares:
Esta gráfica muestra que las líneas pares presentan tintas con una cantidad de hierro notablemente mayor que aquellas que están en las líneas impares. Esto ocurre a pesar de que, como se puede comprobar, visualmente parecen tintas idénticas:

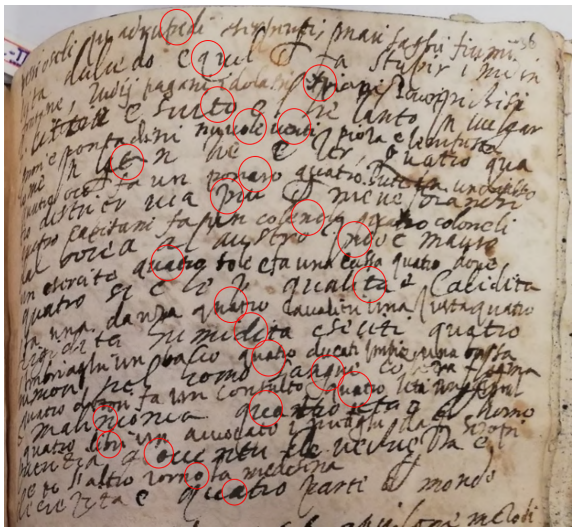

Figura 5.44. Zonas de medida en la página 36R.

Este hecho responde, como se menciona en el artículo [36] publicado hace escasas semanas por este equipo, a la presencia de unas interlíneas (líneas pares) escritas con posterioridad en esta página (podrían asociarse al Cluster 3 generado por GMM). En esto se ve también lo potente que es esta combinación de técnicas.

- **AgglomerativeClustering - WARD linkage**: este método ha generado unos espectros medios casi idénticos a los generados por GMM, teniéndose a lo sumo como diferencia alguna reiteración de clusters. Dado que los clusters son muy parecidos a los generados por el método anteriormente expuesto, se puede considerar que no es interesante hacer el análisis de distribuciones hecho para el método anterior. Aún así, este método de clustering es muy interesante especialmente desde el punto de vista de la filología y la historia del arte porque, al ser un clustering jerárquico, permite la creación de dendogramas, que sirven para ver las relaciones entre tintas.

  Se tiene, por ejemplo, este dendograma básico:

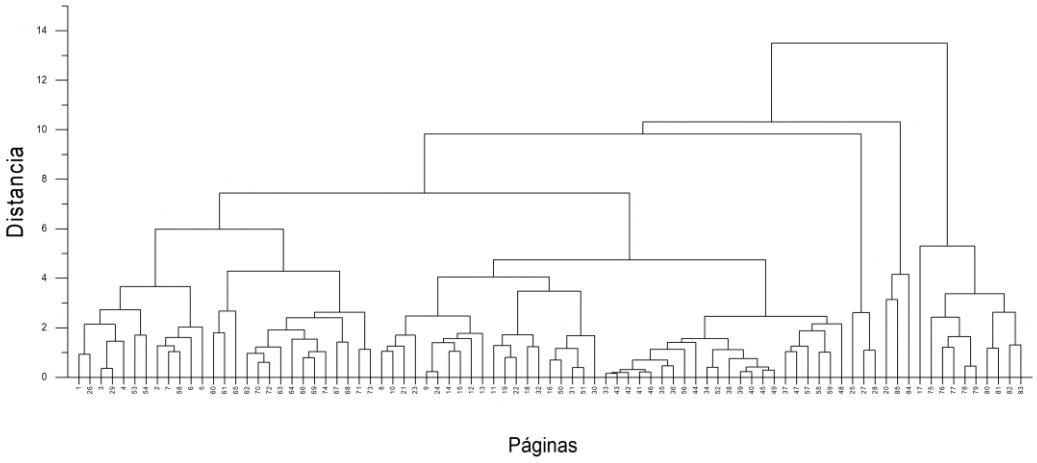

Figura 5.45. Dendograma básico que representa relaciones entre tintas a través de las distintas páginas. [36]

El dendrograma de la Figura 5.45 representa la estructura jerárquica obtenida. En el eje horizontal se sitúan las páginas del manuscrito, mientras que el eje vertical indica la *distancia*, que es la medida de disimilitud en el momento de la fusión de los clústeres. Cada folio comienza como un clúster independiente en la base del gráfico y, a medida que se asciende, se van uniendo en grupos cada vez más amplios. La altura de cada nodo refleja el grado de diferencia existente entre los clústeres que se combinan: fusiones a alturas bajas indican gran similitud, mientras que aquellas que se producen a alturas elevadas señalan grupos más heterogéneos.

Se observan varios conglomerados de folios que se unen a distancias intermedias (en torno a valores de 6-7 en el eje vertical), lo que sugiere la presencia de subconjuntos relativamente homogéneos dentro de la colección. Por otro lado, existen agrupaciones que sólo se fusionan a distancias mucho mayores (superiores a 10), indicando diferencias notables respecto al resto.

- **BIRCH** ha generado una excesiva granularidad, generando 15 clusters muy repetitivos entre sí, siendo algunos de ellos incluso inconsistentes con lo visto en el manuscrito. No obstante, este método ha conseguido aislar un espectro sumamente curioso:

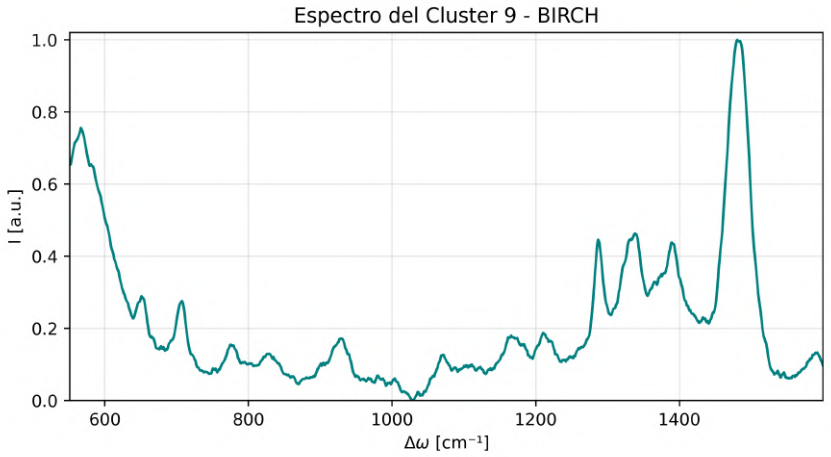

Figura 5.46. Espectro obtenido del Cluster 9 de BIRCH.

Este espectro tiene la siguiente nomenclatura:

36V25_misericordia_r2

Lo que hace tan curioso este espectro es que fue caracterizado previamente por D. Santiago. Esto lo expuso en su conferencia para la International Iron Gall Ink Meeting 2025, en su charla [37].

Como se expone en su conferencia, esta tipología de tinta corresponde a una tinta muy degradada, muy abundante cuando hay exceso de cobre (impurezas de las sales de hierro). En este caso particular apenas aparece por el bajo contenido en cobre de las tintas del manuscrito, lo que permite deducir que se han utilizado unas sales férricas muy puras en este caso.

Para comprobar esto, se hizo una serie de mediciones usando la técnica XRF (Fluorescencia de Rayos-X), ampliamente utilizada para descubrir los elementos químicos presentes en una muestra:

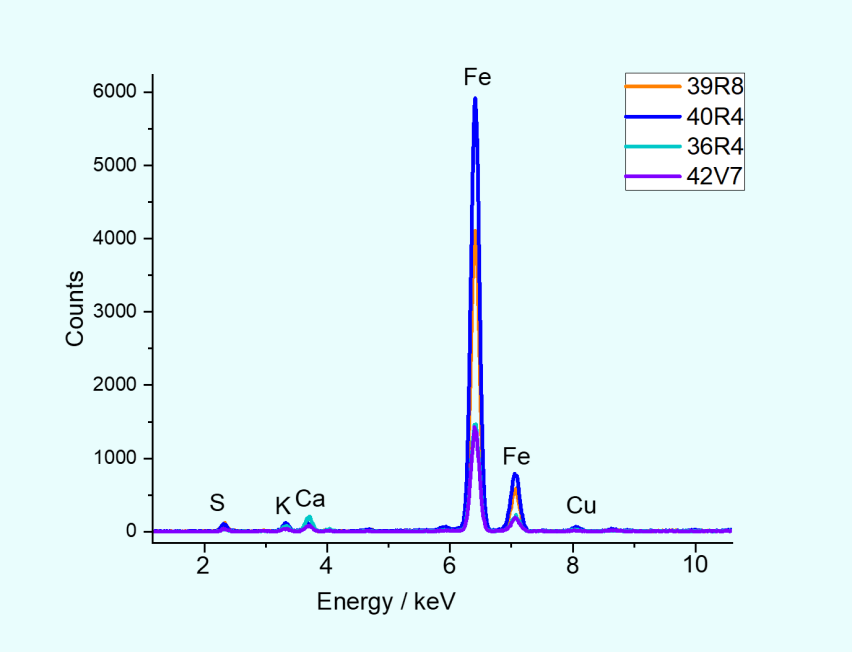

Figura 5.47. Materiales encontrados con XRF en las tintas del manuscrito.

Esto prueba, en efecto, la pureza en hierro de las tintas utilizadas en la elaboración del manuscrito.

Este es un espectro que ha podido ser localizado única y exclusivamente gracias al uso de BIRCH, que es otra cosa que prueba la potencia de este enfoque combinado.

Capítulo 6. CONCLUSIONES

6.1 Conclusiones del trabajo

En este proyecto se han conseguido todos los objetivos propuestos y se ha ahondado en ellos aún más:

1. Se ha conseguido automatizar el preprocesamiento de una gran cantidad de espectros Raman.

2. Se ha creado una aplicación que puede usarse para limpiar de forma manual, intuitiva y rápida espectros Raman sueltos, lo cual puede llegar a ser una herramienta muy útil para los investigadores de este campo.

3. Se ha conseguido crear un proceso replicable para analizar a escala masiva espectros Raman de tintas ferrogálicas de forma automática, si bien es absolutamente necesaria para un correcto uso del mismo la colaboración con especialistas en el campo de estudio. Este método consiste en la obtención de componentes con el algoritmo NMF y la posterior aplicación de métodos de clustering como GMM.

4. La bondad de los resultados sirve para demostrar lo útil que puede resultar la combinación de espectroscopía y Machine Learning en campos tan inesperados como puede ser la filología. Esto abre potencialmente un nuevo campo de investigación y la posibilidad de mejorar y acelerar todos los estudios de este tipo.

5. El manuscrito analizado presenta una dificultad técnica notable, ya que únicamente contiene tinta ferrogálica, sin iluminaciones, pigmentos ni otros materiales que aporten señales diferenciadas. Lo que se han analizado han sido tintas creadas por la misma persona y utilizando en principio la misma receta en todos los casos, conque lo que se ha tenido que detectar es una serie de pequeñas diferencias en espectros prácticamente iguales. En este contexto, los resultados obtenidos adquieren mayor relevancia, al haberse alcanzado en condiciones particularmente complicadas para la caracterización espectroscópica. Es aquí donde se observa que la toma de datos ha sido excelente, pues con unos datos peores la obtención de estos resultados no habría sido posible de ninguna manera.

Como se expone al final del artículo anteriormente mencionado [36], el Dr. Sánchez-Cortés llevó a cabo un análisis muy exhaustivo con los métodos usados tradicionalmente y obtener los resultados expuestos le llevó semanas.

El tiempo de ejecución del programa que desarrolla todo el trabajo expuesto en la sección *Resultados de NMF* (una vez optimizado para este caso particular) ha tomado este tiempo:

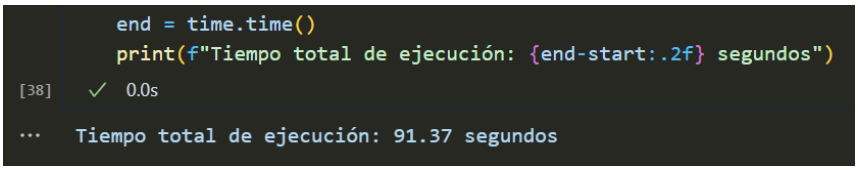

Figura 6.1. Tiempo de ejecución del programa.

Este procedimiento supone una mejora muy grande obteniendo unos resultados muy parecidos a los comentados en el artículo. Los resultados dependen, aún así, de la calidad de los espectros tomados y, en última instancia, la interpretación de estos depende de una valoración experta. Este trabajo muestra que el método presentado es una opción muy buena si se desea reducir tiempos para obtener unos buenos resultados a analizar o incluso establecer una solución en la que basarse de cara a un análisis con métodos más tradicionales.

Este método de trabajo permite también descifrar los componentes de las tintas que están ocultos a simple vista (ácido gálico, oxalato, sílice...), trabajando directamente con el conjunto de datos en crudo y sin tener que analizar los espectros de forma individual para determinar si descartar o no uno, lo cual ahorra también mucho tiempo y carga de trabajo.

6.2 Conclusiones personales

El desarrollo de este proyecto me ha permitido profundizar en un ámbito multidisciplinar en el que convergen la espectroscopía, la filología y las técnicas de análisis de datos. La experiencia ha sido muy enriquecedora, no solo por el reto técnico que ha supuesto, sino también por la posibilidad de aplicar herramientas de Machine Learning a un campo tan singular como el estudio de manuscritos históricos.

Considero especialmente relevante que la aplicación de este tipo de metodologías permite liberar a los investigadores de las tareas más repetitivas y costosas en tiempo, de modo que puedan centrarse en lo que resulta verdaderamente valioso para ellos: el análisis de los resultados y la búsqueda de relaciones dentro del manuscrito. La combinación de la potencia de la espectroscopía Raman con las capacidades de los métodos de Big Data abre nuevas oportunidades, haciendo posible una exploración más ágil, precisa y profunda de este tipo de documentos.

En lo personal, este trabajo me ha permitido aprender a integrar diferentes disciplinas y a valorar la importancia de la colaboración entre especialistas de campos distintos. Además, me ha permitido ver lo importantes que resultarán las herramientas de Machine Learning como herramienta de apoyo fundamental en la investigación humanística y científica, aportando nuevas perspectivas y acelerando procesos que de otro modo resultarían demasiado costosos en tiempo.

Capítulo 7. FUTURAS LÍNEAS DE TRABAJO

El carácter pionero de este proyecto abre diversas posibilidades de investigación que pueden explorarse en trabajos posteriores:

- **Ampliación del corpus de estudio:** aplicar la metodología a otros manuscritos históricos con diferentes características materiales (presencia de pigmentos, iluminaciones, tintas diversas) y comparar resultados entre documentos de distintas épocas o procedencias.
- **Integración de nuevas técnicas espectroscópicas:** complementar el análisis con otras técnicas como FORS (Espectroscopía de Reflectancia de Fibra Óptica), con el fin de desarrollar flujos de trabajo multimodales que combinen espectros de distintas fuentes.
- **Construcción de una base de datos de referencia:** desarrollar un repositorio amplio y estandarizado de espectros Raman procedentes de manuscritos históricos, que sirva como punto de partida para la aplicación de modelos avanzados de *deep learning* para obtener clasificaciones automáticas de compuestos como el ácido gálico o los sulfatos.
- **Herramientas interactivas para investigadores:** mejorar la aplicación desarrollada con funcionalidades de visualización dinámica, selección de regiones del manuscrito y plataformas colaborativas para la anotación conjunta de resultados.
- **Validación y estandarización:** elaborar protocolos replicables y conjuntos de datos de referencia que permitan a otros equipos aplicar y validar esta metodología, contribuyendo a la creación de benchmarks públicos en el campo.

Bibliografía

[1] P. Vandenabeele, H. G. Edwards y L. Moens, «A decade of Raman spectroscopy in art and archaeology,» *Chemical reviews*, vol. 107, n.º 3, págs. 675-686, 2007.

[2] V. Hayez, S. Denoel, Z. Genadry y B. Gilbert, «Identification of pigments on a 16th century Persian manuscript by micro-Raman spectroscopy,» *Journal of Raman Spectroscopy*, vol. 35, n.º 8-9, págs. 781-785, 2004.

[3] R. J. Clark, «Raman microscopy: application to the identification of pigments on medieval manuscripts,» *Chemical Society Reviews*, vol. 24, n.º 3, págs. 187-196, 1995.

[4] F. Ospitali, D. C. Smith y M. Lorblanchet, «Preliminary investigations by Raman microscopy of prehistoric pigments in the wall-painted cave at Roucadour, Quercy, France,» *Journal of Raman Spectroscopy*, vol. 37, n.º 10, págs. 1063-1071, 2006. DOI: https://doi.org/10.1002/jrs.1611. eprint: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/jrs.1611. dirección: https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/jrs.1611.

[5] A. S. Lee, P. J. Mahon y D. C. Creagh, «Raman analysis of iron gall inks on parchment,» *Vibrational Spectroscopy*, vol. 41, n.º 2, págs. 170-175, 2006.

[6] A. S. Lee, V. Otieno-Alego y D. C. Creagh, «Identification of iron-gall inks with near-infrared Raman microspectroscopy,» *Journal of Raman Spectroscopy: An International Journal for Original Work in all Aspects of Raman Spectroscopy, Including Higher Order Processes, and also Brillouin and Rayleigh Scattering*, vol. 39, n.º 8, págs. 1079-1084, 2008.

[7] A. Espina, S. Sanchez-Cortes y Z. Jurašeková, «Vibrational study (Raman, SERS, and IR) of plant gallnut polyphenols related to the fabrication of iron gall inks,» *Molecules*, vol. 27, n.º 1, pág. 279, 2022.

[8] E. Vassiou, D. Lazidou, E. Kampasakali, E. Pavlidou y J. Stratis, «Iron gall ink from historical recipes on organic substrates and their study before and after accelerated ageing with μ-RAMAN Spectroscopy and SEM-EDS,» *Journal of Cultural Heritage*, vol. 66, págs. 584-592, 2024.

[9] R. Bro y A. K. Smilde, «Principal component analysis,» *Analytical methods*, vol. 6, n.º 9, págs. 2812-2831, 2014.

[10] Y.-X. Wang e Y.-J. Zhang, «Nonnegative matrix factorization: A comprehensive review,» *IEEE Transactions on knowledge and data engineering*, vol. 25, n.º 6, págs. 1336-1353, 2012.

[11] B. Melit Devassy, S. George y P. Nussbaum, «Unsupervised clustering of hyperspectral paper data using t-SNE,» *Journal of imaging*, vol. 6, n.º 5, pág. 29, 2020.

[12] T. Sainburg, L. McInnes y T. Q. Gentner, «Parametric UMAP embeddings for representation and semisupervised learning,» *Neural Computation*, vol. 33, n.º 11, págs. 2881-2907, 2021.

[13] X. Chen, J. Shen, C. Liu et al., «Applications of data characteristic AI-assisted raman spectroscopy in pathological classification,» *Analytical Chemistry*, vol. 96, n.º 16, págs. 6158-6169, 2024.

[14] Official UMAP webpage: https://umap-learn.readthedocs.io/en/latest/parameters.html.

[15] J. MacQueen, «Some methods for classification and analysis of multivariate observations,» en *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, University of California press, vol. 5, 1967, págs. 281-298.

[16] C. M. Bishop y N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4.

[17] A. Ng, M. Jordan e Y. Weiss, «On spectral clustering: Analysis and an algorithm,» *Advances in neural information processing systems*, vol. 14, 2001.

[18] T. Zhang, R. Ramakrishnan y M. Livny, «BIRCH: an efficient data clustering method for very large databases,» *ACM sigmod record*, vol. 25, n.º 2, págs. 103-114, 1996.

[19] B. J. Frey y D. Dueck, «Clustering by passing messages between data points,» *science*, vol. 315, n.º 5814, págs. 972-976, 2007.

[20] L. McInnes, J. Healy, S. Astels et al., «hdbscan: Hierarchical density based clustering.,» *J. Open Source Softw.*, vol. 2, n.º 11, pág. 205, 2017.

[21] M. Ankerst, M. M. Breunig, H.-P. Kriegel y J. Sander, «OPTICS: Ordering points to identify the clustering structure,» *ACM Sigmod record*, vol. 28, n.º 2, págs. 49-60, 1999.

[22] P. J. Rousseeuw, «Silhouettes: a graphical aid to the interpretation and validation of cluster analysis,» *Journal of computational and applied mathematics*, vol. 20, págs. 53-65, 1987.

[23] T. Caliński y J. Harabasz, «A dendrite method for cluster analysis,» *Communications in Statistics-theory and Methods*, vol. 3, n.º 1, págs. 1-27, 1974.

[24] D. L. Davies y D. W. Bouldin, «A cluster separation measure,» *IEEE transactions on pattern analysis and machine intelligence*, n.º 2, págs. 224-227, 1979.

[25] C. Remazeilles, V. Quillet y J. Bernard, «FTIR techniques applied to iron gall inked damaged paper,» en *proceedings of the 15th world conference on non-destructive testing*, 2000.

[26] V. Corregidor, R. Viegas, L. M. Ferreira y L. C. Alves, «Study of iron gall inks, ingredients and paper composition using non-destructive techniques,» *Heritage*, vol. 2, n.º 4, págs. 2691-2703, 2019.

[27] A. B. López-Baldomero, M. Buzzelli, F. Moronta-Montero, M. Á. Martínez-Domingo y E. M. Valero, «Ink classification in historical documents using hyperspectral imaging and machine learning methods,» *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, vol. 335, pág. 125 916, 2025.

[28] J. Havermans, H. A. Aziz y H. Scholten, «Non destructive detection of iron-gall inks by means of multispectral imaging part 2: application on original objects affected with iron-gall-Ink corrosion,» 2003.

[29] Wikipedia: https://es.wikipedia.org/wiki/Stefanello_Bottarga.

[30] J. Otero y V. Cano, *Espectroscopía Raman: Fundamento y aplicaciones*, ago. de 2015. DOI: 10.13140/RG.2.1.5015.5362.

[31] N. Coca-Lopez, «An intuitive approach for spike removal in Raman spectra based on peaks' prominence and width,» *Analytica Chimica Acta*, 2024.

[32] https://es.wikipedia.org/wiki/Recto_y_verso.

[33] J. Kolar, A. Štolfa, M. Strlič et al., «Historical iron gall ink containing documents — Properties affecting their condition,» *Analytica Chimica Acta*, vol. 555, n.º 1, págs. 167-174, 2006, ISSN: 0003-2670. DOI: https://doi.org/10.1016/j.aca.2005.08.073. dirección: https://www.sciencedirect.com/science/article/pii/S0003267005014765.

[34] A. del Castillo-Rueda y P. Khosravi-Shahi, «The role of iron in the interaction between host and pathogen,» *Medicina Clinica*, vol. 134, n.º 10, págs. 452-456, 2009.

[35] M. Hanesch, «Raman spectroscopy of iron oxides and (oxy) hydroxides at low laser power and possible applications in environmental magnetic studies,» *Geophysical Journal International*, vol. 177, n.º 3, págs. 941-948, 2009.

[36] M. d. V. Ojeda-Calvo, F. J. Gómez-Fernández, A. Crespo y S. Sánchez-Cortés, «La espectroscopía Raman como herramienta de la filología: análisis del manuscrito II/1391 (2) de la colección Gondomar atribuido a Botarga,» *Avisos. Noticias de la Real Biblioteca*, vol. 31, n.º 105, págs. 4-32, 2025, ISSN: 1578-8334. dirección: https://avisos.realbiblioteca.es.

[37] S. Sanchez-Cortés, «Raman and Surface-Enhanced Raman scattering analysis of iron gall inks employed in historic manuscripts: Effect of aging on the spectral markers,» en *The International Iron Gall Ink Meeting 2025: Towards Sustainable Preservation*, 2025.